

Convergence of Fundamental Limitations in Feedback Communication, Estimation, and Feedback Control over Gaussian Channels

Jialing Liu and Nicola Elia

Abstract

In this paper, we establish the connections of the fundamental limitations in feedback communication, estimation, and feedback control over Gaussian channels, from a unifying perspective for information, estimation, and control. The optimal feedback communication system over a Gaussian necessarily employs the Kalman filter (KF) algorithm, and hence can be transformed into an estimation system and a feedback control system over the same channel. This follows that the information rate of the communication system is alternatively given by the decay rate of the Cramer-Rao bound (CRB) of the estimation system and by the Bode integral (BI) of the control system. Furthermore, the optimal tradeoff between the channel input power and information rate in feedback communication is alternatively characterized by the optimal tradeoff between the (causal) one-step prediction mean-square error (MSE) and (anti-causal) smoothing MSE (of an appropriate form) in estimation, and by the optimal tradeoff between the regulated output variance with causal feedback and the disturbance rejection measure (BI or degree of anti-causality) in feedback control. All these optimal tradeoffs have an interpretation as the tradeoff between causality and anti-causality. Utilizing and motivated by these relations, we provide several new results regarding the feedback codes and information theoretic characterization of KF. Finally, the extension of the finite-horizon results to infinite horizon is briefly discussed under specific dimension assumptions (the asymptotic feedback capacity problem is left open in this paper).

Keywords: Fundamental limitations; Gaussian channels with memory; confluence of feedback communication, estimation, and feedback control; Kalman filtering (KF); minimum mean-square error (MMSE); Bode integral (BI); smoothing, filtering, and prediction; causality versus anti-causality; Cover-Pombra coding structure; Schalkwijk-Kailath scheme; cheap control

I. INTRODUCTION

Communication systems in which the transmitters have access to noiseless feedback of channel outputs have been widely studied. The *fundamental limitations* in these systems, i.e. the feedback capacities, and the capacity-achieving codes, have been a central focus in the information theoretic literature. As one of the most important case, the single-input single-output Gaussian channels with noiseless feedback have attracted considerable attention; see [1]–[15] and references therein for the capacity characterization and coding scheme design for these channels. There exist different approaches in addressing the fundamental limitations for such channels, categorized roughly (by no means strict as the approaches are intrinsically related) as follows: 1) *Estimation theory related approaches*, which utilizes concepts such as maximum likelihood (ML) or minimum mean-square error (MMSE) estimates in constructing the coding schemes (cf. e.g. [1], [2], [4], [5], [7], [16]); 2) *Information theoretic approaches*,

This research was supported by NSF under Grant ECS-0093950. The material in this paper was presented in part at the 43rd Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, September 2005, and the 45th IEEE Conference on Decision and Control, San Diego, CA, December 2006.

J. Liu was with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA. He is now with Motorola Inc., Libertyville, IL 60048 USA (e-mail: jialingliu@motorola.com).

N. Elia is with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: nelia@iastate.edu).

most notably the Cover-Pombra formulation based on the asymptotic equipartition (AEP) property and the mutual information between the message and the channel outputs (cf. e.g. [6], [9], [15]), and the directed information formulation based on the input-output characterization of the channels ¹ (cf. e.g. [11], [17], [18]); and 3) *Control theory related approaches*, which regards the feedback communication problems as optimal control problems (cf. e.g. [3], [11]–[14]).

In particular, Schalkwijk and Kailath [1], [2] proposed the Schalkwijk-Kailath (SK) codes for additive white Gaussian noise (AWGN) channels, achieving the asymptotic feedback capacity (i.e. the infinite-horizon feedback capacity, denoted C_∞ , which is the highest information rate over the time spans between 0 and infinity, subject to an average power constraint) and greatly reduce the coding complexity and coding delay. The SK codes were suggested by the Robbins-Monro stochastic approximation and recursive ML algorithm which have an *estimation* theoretic flavor. Along the line of [1], [2], Butman, Ozarow, and numerous other researchers have proposed extensions of the SK codes to Gaussian feedback channels with memory and obtained tight capacity bounds, see e.g. [4], [5], [7].

Cover and Pombra [6] introduced a general coding structure (called the Cover-Pombra structure, or the CP structure for short) to achieve the finite-horizon feedback capacity (denoted C_T , the highest information rate over the time span between 0 and T subject to an average power constraint) for Gaussian channels with memory, based on classical *information* theoretic concepts. Their development builds on the mutual information between the message and the channel outputs (hence circumventing the causality issue pointed out by Massey [17] without appealing to directed information) and AEP for arbitrary Gaussian processes. The CP structure was initially regarded to have prohibitive computation complexity if the coding length $(T + 1)$ is large (see, however, Section IV-A for more detailed discussion), and efforts have been made to reduce the complexity and to refine the CP structure. By exploiting the special properties of a moving-average Gaussian channel with feedback, Ordentlich [9] discovered the finite rank property of the innovations in the CP structure, which reduces the computation complexity. Shahar-Doron and Feder [10] reformulated the CP structure along this direction, and obtained an SK-based coding scheme to achieve C_T with reduced computation complexity. Furthermore, utilizing the CP structure as a starting point, Kim [15] proved that a closed-form expression ² of the asymptotic capacity C_∞ for an first-order moving-average Gaussian channel with feedback, and obtained an SK-based coding scheme to achieve C_∞ . This is the first Gaussian channel with memory (except for the degenerated case of AWGN channel) ³ that has an established asymptotic feedback capacity and available capacity-achieving codes, to the best of our knowledge. On the other hand, Vandenberghe *et al* [19] showed that the computation of C_T based on the CP structure can be reformulated as a convex optimization problem.

Tatikonda and Mitter [11], [18] provided an extensive study of feedback communication systems and their capacities. They extended the notion of *directed information* proposed in [17] and proved that its supremum equals the operational capacity; reformulated the problem of computing C_T as a *stochastic control optimization problem*; and proposed a dynamic programming based solution and characterized the sufficient statistics required for encoding and decoding. This idea was further explored in [12] by Yang *et al*, which uncovered the Markov property of the optimal input distributions for Gaussian channels with memory, established a class of refined, finite-

¹The directed information in feedback communication systems may be viewed as the causal counterpart of mutual information used in communication systems without feedback, the supremum of which (under applicable constraints, if any) is the capacity. See also Appendix II.

²This expression was initially identified by Elia [14] and Yang *et al* [12] and has been conjectured to be C_∞ ; however, a rigorous proof was not available until Kim [15].

³By Gaussian channels with memory, researchers normally refer to frequency-selective Gaussian channels, including Gaussian channels with inter-symbol interference (ISI) and channels with colored Gaussian noise, a convention also adopted in this paper (although some other Gaussian channels may also have memory). The Gaussian channels with memory may sometimes be referred to as general Gaussian channels (in contrast to the specific AWGN channels), or even simply as Gaussian channels.

dimensional optimal input distributions, and eventually reduced the finite-horizon stochastic control optimization problem to a manageable size (with complexity $O(T)$). Moreover, under a *stationarity conjecture* that C_∞ can be achieved by a stationary input process, C_∞ is given by the solution of a finite-dimensional optimization problem. This is the first computationally efficient ⁴ method to calculate the feedback capacity in infinite horizon for general Gaussian channels. A Kalman filter (KF) was used in [12] to generate the sufficient statistics of the output feedback.

Omura [3] identified a stochastic *optimal control problem* for feedback communication systems. Omura showed that the solution to the control problem is optimal for AWGN channels in the sense of achieving the capacity; however, how this approach might be extended to achieve the capacities of more general channels remained to be seen ⁵. Sahai and Mitter [13], [20] investigated the problem of tracking unstable sources over a channel and introduced the notion of *anytime capacity* to capture the fundamental limitations in that problem, which again reveals connections between communication and control and brings various new insights to feedback communication problems. Furthermore, Elia [14] established the *equivalence* between reliable communication and stabilization over Gaussian channels with memory, showed that the achievable transmission rate is given by the Bode sensitivity integral of the associated control system, and presented an optimization problem based on robust control to compute lower bounds of C_∞ . These lower bounds can be achieved by generalized SK codes that have an interpretation of tracking unstable sources over Gaussian channels. For a time-varying fading AWGN channel whose fade is modelled as a Markov process with channel output feedback and channel state information (CSI), a *control-oriented* coding scheme multiplexing across multiple subsystems according to CSI was constructed by Liu *et al* [21] to achieve the ergodic capacity, and it is shown to be an extension of the SK codes to time-varying channels with appropriate channel state information. For a recent survey of various topics on feedback communication, see e.g. [15], [22] and references therein.

As we have seen, different approaches have been shown useful in addressing the Gaussian feedback communication problem. This paper attempts to present a converging point: We study the Gaussian channels with feedback from a perspective that unifies information, estimation, and control, which encompasses many of the existing approaches scattered in the literature. We demonstrate that the feedback communication problem over a Gaussian channel can be reformulated as an optimal estimation problem or an optimal control problem. In fact, we show that the existing coding structures either necessarily contain Kalman filters or are reformulations of Kalman filters: The CP structure necessitates a KF in order to be optimal, the SK code can be easily obtained or extended by transforming a KF, and the control-oriented schemes can be derived from a KF by the duality between control and estimation [23]. As a result, *the fundamental limitations in feedback communication, estimation, and feedback control coincide*.

Particularly, the achievable rate of the feedback communication system is alternatively given by the decay rate of the Cramer-Rao bound (CRB) for the associated estimation system as well as the Bode integral (BI) of the associated control system. In addition, the fundamental limitations in terms of the *optimal tradeoffs in feedback communication, estimation, and feedback control coincide, all of which may be interpreted as the tradeoff between causality and anti-causality*. In feedback communication, this fundamental limitation is the optimal tradeoff between the input power and information rate. Alternatively in the associated estimation system, it can be characterized by the optimal tradeoff between the (causal) one-step prediction and (anti-causal) smoothing, or in the associated control system by the optimal tradeoff between the variance of a regulated output (generated using causal feedback) and the BI (or degree of anti-causality or instability). That is, the optimal pairs (P, R) , $(\text{PMMSE}_T, (\log \det \text{MMSE}_T^{-1})/(2T+2))$,

⁴Here we do not mean that their optimization problem is convex. The computation complexity associated with the optimization problem is determined mainly on the channel order which does not grow to infinity as the time horizon increases to infinity.

⁵Rather than showing the feedback capacity problem can be posed as a control problem as Tatikonda and Mitter did, Omura formulated the control problem to minimize MMSE. Whether this may yield information theoretic optimality was not explored by Omura [3] except for the AWGN case. Later works such as [11], [12], [14], [20] and the present paper have established results on the intrinsic relationship between communication and control within a more general framework.

and $(P_u, \log DI)$ correspond to each other, where P is the average channel input power and R is the average information rate in the communication system; $PMMSE_T$ is the time average of the *one-step prediction MMSE* of the to-be-estimated process in the estimation system and $MMSE_T$ is the *anti-causal smoothing MMSE* of the initial state of the process; P_u is the variance of the regulated output u (i.e. control performance measure) in the control system and DI is the *degree of instability* of the open-loop system defined as the product of open-loop unstable eigenvalues and is equal to the *Bode sensitivity integral* (i.e. disturbance rejection measure). Here the tradeoffs mean that if one wishes to keep the first element in the pair small (such as low channel input power), the other element cannot be made arbitrarily large. See Sec. VII-C for more precise descriptions. We call DI the degree of anti-causality since it is associated with right-half plane (RHP) poles. Note that references exist in addressing various aspects of fundamental limits; for an incomplete list, Van Trees [24] (pp. 501-511), de Bruijn, and Guo *et al* [25] (and therein references and subsequent works) discussed filtering versus smoothing as well as their relation to entropy and mutual information, Feng *et al* [26] examined the KF MMSE performance related to information theoretic measures, Iglesias and coauthors [27], [28] studied BI and its information theoretic interpretation, Seron *et al* [29] presented connections of the fundamental limitations between control and filtering, Martins and Dahleh [30], [31] studied BI and entropy rates for systems over communication channels. See also [11], [20] and more discussions in Sec. VII-A.

Utilizing or motivated by the above mentioned equivalence relationship, we provide 1) New refinements to the Cover-Pombra capacity-achieving coding structure, including the complete characterization of the feedback generator; the necessity of KF in the CP structure; the orthogonality between future channel inputs and past channel outputs; the Gauss-Markov property of the transformed channel outputs; and the finite-dimensionality of the optimal message-carrying inputs. 2) Simple equivalence between generalized Schalkwijk-Kailath codes and the KF, which yields a convenient way to obtain a feedback communication scheme from an estimation problem. 3) *Information theoretic characterization of KF*; that is, the KF is not only a device to provide sufficient statistics (which was shown in [12]), but also a device to ensure the power efficiency and to recover the message optimally. 4) *The necessity of MMSE estimation* in feedback communication problems over general additive noise channels with an average power constraint. Our results 1) - 3) hold for AWGN channels with intersymbol interference (ISI) where the ISI is modelled as a stable and minimum-phase FDLTI system; through the equivalence shown in [11], [12], this channel is equivalent to a colored Gaussian channel with a rational noise power spectrum (which is assumed in a number of references) and without ISI. The above results are mainly derived in the finite horizon, but we also show that the KF converges to a steady state as time goes to infinity, and the equivalence holds in the steady state system as well. Note that, however, the infinite-horizon feedback capacity (or the stationary feedback capacity) problem is left open in this paper ⁶.

This paper is organized as follows. In Section II, a motivating example of feedback communication over an AWGN channel is presented. In Section III, we describe the general Gaussian channel models. We then introduce the feedback capacity in finite horizon and the CP structure in Section IV. In Section V, we consider a general coding structure in finite-horizon which is closely related to the CP structure but allows us to easily see the necessity of the KF algorithm in feedback communication. The presence of the KF links the feedback communication problem to an estimation problem and a control problem as shown in Section VI, and hence we rewrite the information rate and input power in terms of estimation theory quantities and control theory quantities and explore the connections; see Section VII. More necessary conditions for the optimality of the coding structure are proposed in Section VIII. Sections V to VIII are focused on finite horizon. In Section IX, we extend the horizon to infinity and characterize the steady-state behavior.

Notations: We use underlines to specify vectors, and use boldface to specify matrices. To ease the reading, all

⁶We note that Kim in [32] and further in [33] claims the stationary conjecture is verified. This leads to that stationary feedback capacity equals the asymptotic feedback capacity.

vectors in this paper are column vectors. We represent transpose by $'$. We represent time indices by subscripts, such as y_t . We denote by \underline{y}^T the collection $\{y_0, y_1, \dots, y_T\}$, and $\{y_t\}$ the sequence $\{y_t\}_{t=0}^\infty$. We assume that the starting time of all processes is 0, consistent with the convention in dynamical systems but different from the information theory literature. We use $h(X)$ for the differential entropy of the random variable X . For a random vector \underline{y}^T , we denote its covariance matrix as $\mathbf{K}_{\underline{y}}^{(T)}$. The norm $\|\underline{y}\|$ is the Euclidean norm of the vector. We denote $\mathcal{T}_{xy}(z)$ as the transfer function from x to y . As a linear input-output relation (linear system) $\mathcal{Z}(z)$ can be alternatively captured by a matrix, we represent the matrix associated with linear system $\mathcal{Z}(z)$ by $\mathbf{Z}(z)$ (boldface script Z). We denote “defined to be” as “:=”. We use $(\mathbf{A}, \mathbf{B}, \mathbf{C}', D)$ to represent system

$$\begin{cases} \underline{x}_{t+1} &= \mathbf{A}\underline{x}_t + \mathbf{B}u_t \\ y_t &= \mathbf{C}'\underline{x}_t + Du_t. \end{cases} \quad (1)$$

Finally, in this paper, by “capacity” we refer to the feedback capacity, if not specified otherwise.

II. MOTIVATING EXAMPLE: FEEDBACK CAPACITY AND OPTIMAL SCHEMES FOR AN AWGN CHANNEL

To help the reader understand the intuition behind our study, we present a simple example over an AWGN channel before we go into the Gaussian channels with memory. Below, we introduce a simple KF system (see Fig. 1 (a)), followed by a straightforward rewrite of it (see Fig. 1 (b)), which now has an interpretation as a feedback communication system. Finally we show that this feedback communication system is optimal as it is equivalent to the optimal SK scheme. It motivates the further exploration of the connections among feedback communication, estimation, and feedback control.

1) *A Kalman Filter Problem:* Consider a standard KF problem for a first-order unstable LTI system with noisy measurements:

$$\text{to-be-estimated system: } \begin{cases} x_{t+1} &= ax_t \\ r_t &= cx_t \\ \bar{y}_t &= r_t + N_t, \end{cases} \quad (2)$$

where x_0 is unknown, $a > 1$ (namely the system is unstable), a and c are known, and $N_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. The KF provides MMSE estimate of $\{x_t\}$ based on the noisy measurement process $\{\bar{y}_t\}$. The (steady-state)⁷ KF is described as (See Fig. 1 (a) for the block diagram)

$$\text{Kalman filter: } \begin{cases} \hat{x}_{t+1} &= a\hat{x}_t + Le_t \\ \hat{r}_t &= c\hat{x}_t \\ e_t &= \bar{y}_t - c\hat{x}_t, \end{cases} \quad (3)$$

where

$$L := \frac{a\Sigma c}{1 + c^2\Sigma} \quad (4)$$

is the asymptotic *Kalman filter gain*, and Σ is the asymptotic error covariance for \hat{x}_t (i.e. $\Sigma = \lim_{t \rightarrow \infty} \mathbf{E}(x_t - \hat{x}_t)(x_t - \hat{x}_t)'$), which is the positive solution to the discrete-time algebraic Riccati equation (DARE)

$$\Sigma = a^2\Sigma - \frac{a^2c^2\Sigma^2}{1 + c^2\Sigma}. \quad (5)$$

Solving the DARE, we obtain

$$\Sigma = \frac{a^2 - 1}{c^2}, \quad L = \frac{a^2 - 1}{ac}. \quad (6)$$

⁷Though $\{\bar{y}_t\}$ is neither stationary nor even asymptotically stationary, a time-varying or time-invariant (steady-state) KF can be built to guarantee *bounded* error covariance for estimating x_t , and the difference between the time-varying one and time-invariant one vanishes as time increases, as pointed out in Chapter 14 of [23].

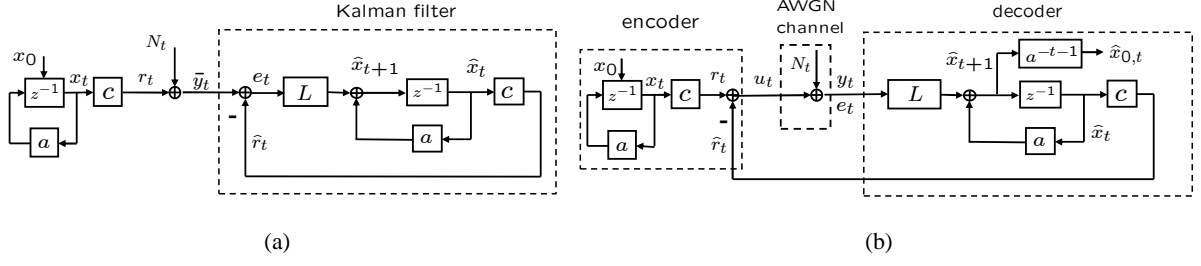


Fig. 1. (a) A KF problem. (b) A KF-based coding structure.

2) *KF-based Feedback Communication*: Next, as illustrated in Fig. 1 (b), we introduce a *feedback communication coding scheme over an AWGN channel* by slightly changing the KF problem shown in Fig. 1 (a). Rather than closing the loop after the AWGN N_t (i.e. adding $(-\hat{r}_t)$ to \tilde{y}_t), in Fig. 1 (b), the loop is closed before the AWGN N_t (i.e. adding $(-\hat{r}_t)$ to r_t). This does not change anything but the signals between the two adders. As indicated in Fig. 1 (b), one can identify the encoder, the AWGN channel, and the decoder, described in the following for time $t = 0, 1, \dots$.

$$\text{AWGN channel: } y_t = u_t + N_t, \quad (7)$$

where u_t is the channel input, $N_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ is the channel noise, and y_t is the channel output. At time t , the encoder can access \hat{r}_t (generated from \underline{y}^{t-1}) via the noiseless feedback link:

$$\text{encoder dynamics: } \begin{cases} x_{t+1} &= ax_t \\ r_t &= cx_t \\ u_t &= r_t - \hat{r}_t \end{cases} \quad (8)$$

where a and c are encoder design parameters. The *encoding procedure* is: Fix a set of M_T equally likely messages, then equally partition the interval $[-\frac{1}{2}, \frac{1}{2}]$ into M_T sub-intervals, and map the sub-interval centers to the set of M_T messages; this is known to both the transmitter and receiver *a priori*. To transmit, let $x_0 := W$, the sub-interval center representing the to-be-transmitted message. In other words, the initial condition (at time 0) of the transmitter is the to-be-transmitted message.

$$\text{decoder dynamics: } \begin{cases} \hat{x}_{t+1} &= a\hat{x}_t + Ly_t \\ \hat{r}_t &= c\hat{x}_t \\ \hat{x}_{0,t} &= a^{-t-1}\hat{x}_{t+1}, \end{cases} \quad (9)$$

and the *decoding procedure* is to simply map $\hat{x}_{0,T}$ into the closest sub-interval center. (Note that in Fig. 1 (b), $y_t = e_t$.)

The *objective* of the feedback communication problem is to, under an average channel input power constraint

$$\frac{1}{T+1} \mathbf{E} \|\underline{u}^T\|^2 \leq \mathcal{P} \text{ or } \lim_{T \rightarrow \infty} \frac{1}{T+1} \mathbf{E} \|\underline{u}^T\|^2 \leq \mathcal{P} \quad (10)$$

with $\mathcal{P} > 0$ being the power budget, achieve

$$C_{fb}(\mathcal{P}) = C_{nf}(\mathcal{P}) = \frac{1}{2} \log(1 + \mathcal{P}), \quad (11)$$

where $C_{fb}(\mathcal{P})$ is the feedback capacity and $C_{nf}(\mathcal{P})$ is the non-feedback capacity in either the finite horizon (time 0 to T) or infinite horizon (time 0 to ∞). To attain this objective, one can fix any coding length $(T + 1)$ and any $\epsilon > 0$ (where ϵ is an arbitrarily small slack from the capacity C_{fb}). Then let $a := \sqrt{1 + \mathcal{P}}$, $c \neq 0$ be arbitrary, $M_T := a^{(T+1)(1-\epsilon)}$, and follow the above-described encoding/decoding dynamics/procedures. It can be shown that this communication scheme can transmit any message out of totally M_T messages with vanishing probability of error as $T \rightarrow \infty$ while satisfying the power constraint (10). Instead of proving the optimality directly, we may alternatively show that the coding scheme in Fig. 1(b) is a simple reformulation of the well-known SK coding scheme that has been shown to achieve the feedback capacity of the AWGN channel. To this aim, a slight variation of the original SK scheme proposed in [2] is illustrated in Fig. 2⁸. In this figure, one can identify the encoder, AWGN channel, decoder, and the feedback link with one-step delay.

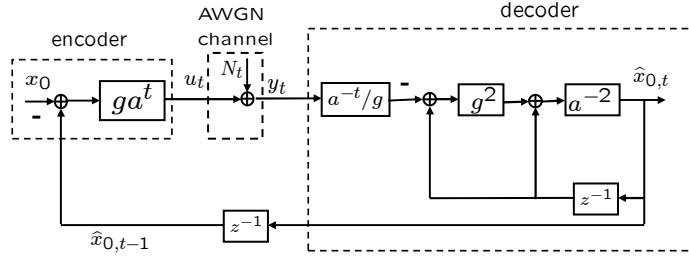


Fig. 2. The SK coding scheme.

To see the connection between the two coding schemes, note that in the SK scheme, it holds that

$$\begin{aligned} u_t &= ga^t(\hat{x}_{0,t-1} - x_0) \\ \hat{x}_{0,t} &= \frac{g^2 + 1}{a^2} \hat{x}_{0,t-1} - a^{-t-2} gy_t; \end{aligned} \quad (12)$$

and in the KF-based scheme, it holds that

$$\begin{aligned} u_t &= ca^t(x_0 - \hat{x}_{0,t-1}) \\ \hat{x}_{0,t} &= \hat{x}_{0,t-1} + a^{-t-1} Ly_t. \end{aligned} \quad (13)$$

If we define

$$g := \sqrt{a^2 - 1}, \quad c := -g, \quad (14)$$

then both schemes generate identical channel inputs, outputs, and decoder estimates respectively, and hence they are considered as equivalent. The optimal choice of g in the SK coding scheme indeed corresponds to the (optimal) KF gain. Thus, we conclude that *the SK scheme essentially implements the KF algorithm*. In fact, more insights can be obtained from this AWGN example; see Chapter 3 of [22]. These insights can be extended to the case of Gaussian channels with memory, which we now turn to.

III. CHANNEL MODEL

In this section, we briefly describe two Gaussian channel models, namely the colored Gaussian noise channel without ISI and white Gaussian noise channel with ISI.

⁸A few SK-type schemes and their variations are compared in [21]. The variation here performs the same operations every step, as opposed to the scheme in [2] in which the initialization step differs from later steps. See also [14], [34]

A. Colored Gaussian noise channel without ISI

Fig. 3 (a) shows a colored Gaussian noise channel without ISI. At time t , this discrete-time channel is described as

$$\tilde{y}_t = u_t + Z_t, \quad \text{for } t = 0, 1, \dots, \quad (15)$$

where u_t is the channel input, Z_t is the channel noise, and \tilde{y}_t is the channel output. We make the following assumptions: The colored noise $\{Z_t\}$ is the output of a finite-dimensional stable and minimum-phase linear time-invariant (LTI) system $\mathcal{Z}(z)$, driven by a white Gaussian process $\{N_t\}$ with zero mean and unit variance, and $\mathcal{Z}(z)$ is at initial rest. We assume that the LTI system $\mathcal{Z}(z)$ has order (or dimension) m and $\mathcal{Z}(\infty) \neq 0$ (i.e. $\mathcal{Z}(z)$ is proper but non-strictly proper). We further assume, without loss of generality, that $\mathcal{Z}(\infty) = 1$; for cases where $g := \mathcal{Z}(\infty) \neq 1$, we can normalize $\mathcal{Z}(z)$ using a scaling factor $1/g$. Then, the finite dimensionality of $\mathcal{Z}(z)$ implies that $\mathcal{Z}(z)$ admits the following *transfer function representation*

$$\mathcal{Z}(z) = \frac{z^m + f_{m-1}z^{m-1} + \dots + f_1z + f_0}{z^m + (f_{m-1} + g_{m-1})z^{m-1} + \dots + (f_1 + g_1)z + (f_0 + g_0)}, \quad (16)$$

where $\{f_0, \dots, f_{m-1}\}$ and $\{g_0, \dots, g_{m-1}\}$ are such that $\mathcal{Z}(z)$ is stable and minimum phase. Define

$$\begin{aligned} \mathcal{Z}_z(z) &= \frac{z^m + f_{m-1}z^{m-1} + \dots + f_1z + f_0}{z^m} \\ \mathcal{Z}_p(z) &= \frac{z^m + (f_{m-1} + g_{m-1})z^{m-1} + \dots + (f_1 + g_1)z + (f_0 + g_0)}{z^m}. \end{aligned} \quad (17)$$

Then it holds that

$$\mathcal{Z}(z) = \frac{\mathcal{Z}_z(z)}{\mathcal{Z}_p(z)}, \quad (18)$$

that is, $\mathcal{Z}_p(z)$ and $\mathcal{Z}_z(z)$ contain the information about the poles and zeros of $\mathcal{Z}(z)$, respectively. For future reference, we define

$$\begin{aligned} \underline{G}_z &:= [f_{m-1}, \dots, f_0]' \\ \underline{G}_p &:= [f_{m-1} + g_{m-1}, \dots, f_0 + g_0]', \end{aligned} \quad (19)$$

that is, \underline{G}_z' and \underline{G}_p' are the output matrices (vectors) for systems $\mathcal{Z}_z(z)$ and $\mathcal{Z}_p(z)$ (see Appendix I-A for relevant state-space representation concepts).

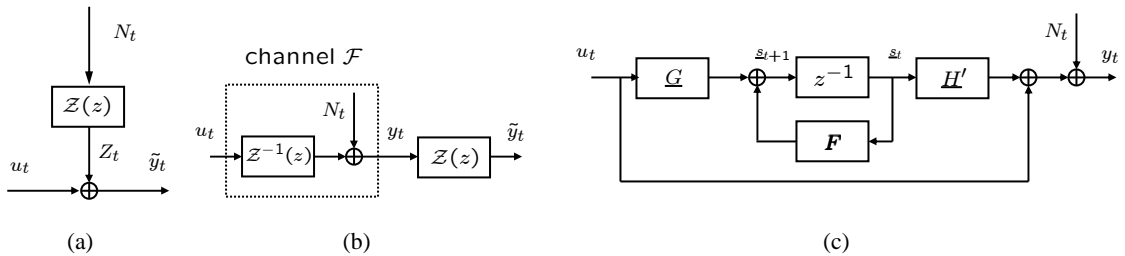


Fig. 3. (a) A colored Gaussian noise channel without ISI. (b) The induced ISI channel with AWGN. (c) State-space realization of channel \mathcal{F} .

We can also represent the input-output relation based on *time-domain operators (matrices)*. For any block size (i.e. coding length) of $(T + 1)$, we may equivalently generate \underline{Z}^T by

$$\underline{Z}^T = \underline{Z}_T \underline{N}^T, \quad (20)$$

where \mathbf{Z}_T is a $(T+1) \times (T+1)$ lower triangular Toeplitz matrix of the impulse response of $\mathcal{Z}(z)$. Since we have assumed that $\mathcal{Z}(\infty) = 1$, the diagonal elements of \mathbf{Z}_T are all 1. Likewise, $\mathbf{Z}_{z,T}$ and $\mathbf{Z}_{p,T}$, the matrix representations of $\mathcal{Z}_z(z)$ and $\mathcal{Z}_p(z)$, are respectively given by

$$\mathbf{Z}_{z,T} := \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ f_{m-1} & 1 & 0 & \cdots & 0 & 0 \\ f_{m-2} & f_{m-1} & 1 & \cdots & 0 & 0 \\ \vdots & & & & & \\ f_0 & f_1 & f_2 & \cdots & 0 & 0 \\ 0 & f_0 & f_1 & \cdots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \cdots & f_{m-1} & 1 \end{bmatrix}$$

$$\mathbf{Z}_{p,T} := \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ f_{m-1} + g_{m-1} & 1 & \cdots & 0 & 0 \\ f_{m-2} + g_{m-2} & f_{m-1} + g_{m-1} & 1 & \cdots & 0 & 0 \\ \vdots & & & & & \\ f_0 + g_0 & f_1 + g_1 & f_2 + g_2 & \cdots & 0 & 0 \\ 0 & f_0 + g_0 & f_1 + g_1 & \cdots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \cdots & f_{m-1} + g_{m-1} & 1 \end{bmatrix}, \quad (21)$$

that is, $\mathbf{Z}_{p,T}$ and $\mathbf{Z}_{z,T}$ are lower triangular, Toeplitz, and banded with bandwidth $(m+1)$, corresponding to causal, LTI, m th order moving-average (MA- m) filters. Therefore, it holds that

$$\mathbf{Z}_T = \mathbf{Z}_{p,T}^{-1} \mathbf{Z}_{z,T} = \mathbf{Z}_{z,T} \mathbf{Z}_{p,T}^{-1}. \quad (22)$$

As a consequence of the above assumptions, $\{Z_t\}$ is asymptotically stationary. Note that there is no loss of generality in assuming that $\mathcal{Z}(z)$ is stable and minimum-phase (cf. Chapter 11, [35]).

B. White Gaussian channel with ISI

The above colored Gaussian channel induces a white Gaussian channel with ISI. More precisely, notice that from (15) and (20), we have

$$\tilde{\mathbf{y}}^T = \mathbf{Z}_T (\mathbf{Z}_T^{-1} \mathbf{u}^T + \mathbf{N}^T), \quad (23)$$

which we identify as a stable and minimum-phase ISI channel with AWGN $\{N_t\}$, see Fig. 3 (b); \mathbf{Z}_T^{-1} is well defined since \mathbf{Z}_T is lower triangular with diagonal elements being $\mathcal{Z}(\infty) = 1$. Here $\mathcal{Z}^{-1}(z)$ is also at initial rest. Note that \mathbf{Z}_T^{-1} is the matrix inverse of \mathbf{Z}_T , equal to the lower-triangular Toeplitz matrix of impulse response of $\mathcal{Z}^{-1}(z)$. For any fixed \mathbf{u}^T and \mathbf{N}^T , (15) and (23) generate the same channel output $\tilde{\mathbf{y}}^T$.⁹

The initial rest assumption on $\mathcal{Z}^{-1}(z)$ can be imposed in practice as follows. First, before a transmission, drive the initial condition (which is enabled by the controllability requirement stated below) of the ISI channel to any desired value that is also known to the receiver a priori. Then, after the transmission, remove the response due to that initial condition at the receiver. Such an assumption is also used in [11], [12].

We can then write the *minimal state-space representation* of $\mathcal{Z}^{-1}(z)$ as $(\mathbf{F}, \mathbf{G}, \mathbf{H}', 1)$, where $\mathbf{F} \in \mathbb{R}^m$ is stable, (\mathbf{F}, \mathbf{G}) is controllable, $(\mathbf{F}, \mathbf{H}')$ is observable, and m is the *dimension* or *order* of $\mathcal{Z}^{-1}(z)$. Let us denote the channel from u to y in Fig. 3(b) as \mathcal{F} , where

$$\mathbf{y}^T := \mathbf{Z}_T^{-1} \mathbf{u}^T + \mathbf{N}^T = \mathbf{Z}_T^{-1} \tilde{\mathbf{y}}^T. \quad (24)$$

⁹More rigorously, the mappings from (u, N) to $\tilde{\mathbf{y}}$ are T -equivalent. For a discussion about systems representations and equivalence between different representations, see Appendix I.

The channel \mathcal{F} is described in *state-space* as

$$\text{channel } \mathcal{F} : \begin{cases} \underline{s}_{t+1} &= \mathbf{F}\underline{s}_t + \underline{G}u_t \\ y_t &= \underline{H}'\underline{s}_t + u_t + N_t, \end{cases} \quad (25)$$

where $\underline{s}_0 = \underline{0}$; see Fig. 3 (c). Notice that channel \mathcal{F} is not essentially different than the channel from u to \tilde{y} , since $\{\underline{y}^t\}$ and $\{\tilde{y}^t\}$ causally determine each other. Without loss of generality, we can choose $(\mathbf{F}, \underline{G}, \underline{H}', 1)$ to have the following observable canonical form:

$$\begin{aligned} \mathbf{F} &:= \left[\begin{array}{c|c} -f_{m-1} & \mathbf{I}_{m-1} \\ \vdots & \\ -f_1 & \\ \hline -f_0 & \underline{0}_{(m-1) \times 1} \end{array} \right], \quad \underline{G} := \begin{bmatrix} g_{m-1} \\ \vdots \\ g_0 \end{bmatrix} \\ \underline{H}' &:= [1 \quad 0 \quad \cdots \quad 0]. \end{aligned} \quad (26)$$

In other words, it holds that $\mathcal{Z}^{-1}(z) = \underline{H}'(z\mathbf{I} - \mathbf{F})^{-1}\underline{G} + 1$. Note that we also have $\underline{G}'_p = \underline{G}' + \underline{G}'_z$ (see (19)).

We concentrate on the case $m \geq 1$; the case that m is 0 (i.e., \mathcal{F} is an AWGN channel) was solved in [1], [2].

IV. THE FEEDBACK CAPACITY IN FINITE-HORIZON AND THE COVER-POMBRA STRUCTURE

A. The CP structure for the colored Gaussian noise channel and finite-horizon capacity

We briefly review the CP coding structure for the colored Gaussian noise channel specified in Section III-A (see [6], [36]). Denote the covariance matrix of the colored Gaussian noise \underline{Z}^T as $\mathbf{K}_{\underline{Z}}^{(T)}$, and let

$$\underline{u}^T := \mathbf{B}_T \underline{Z}^T + \underline{v}^T, \quad (27)$$

where \mathbf{B}_T is a $(T+1) \times (T+1)$ strictly lower triangular matrix, \underline{v}^T is Gaussian with covariance $\mathbf{K}_{\underline{v}}^{(T)} \geq 0$ and is independent of \underline{Z}^T ¹⁰. Now the channel output is

$$\tilde{\underline{y}}^T = \underline{u}^T + \underline{Z}^T = (\mathbf{I} + \mathbf{B}_T)\underline{Z}^T + \underline{v}^T. \quad (28)$$

Then C_T , the *finite-horizon capacity*, is defined as the highest information rate that the CP structure can generate:

$$\begin{aligned} C_T := C_T(\mathcal{P}) &:= \sup \frac{1}{T+1} I(\underline{v}^T; \tilde{\underline{y}}^T) \\ &= \sup \frac{1}{2(T+1)} \log \frac{\det \mathbf{K}_{\tilde{\underline{y}}}^{(T)}}{\det \mathbf{K}_{\underline{Z}}^{(T)}} \\ &= \sup \frac{1}{2(T+1)} \log \frac{\det((\mathbf{I} + \mathbf{B}_T)\mathbf{K}_{\underline{Z}}^{(T)}(\mathbf{I} + \mathbf{B}_T)' + \mathbf{K}_{\underline{v}}^{(T)})}{\det \mathbf{K}_{\underline{Z}}^{(T)}}, \end{aligned} \quad (29)$$

where the supremum is taken over all admissible $\mathbf{K}_{\underline{v}}^{(T)}$ and \mathbf{B}_T satisfying the power constraint

$$P_T := \frac{1}{T+1} \text{tr}(\mathbf{B}_T \mathbf{K}_{\underline{Z}}^{(T)} \mathbf{B}_T' + \mathbf{K}_{\underline{v}}^{(T)}) \leq \mathcal{P}. \quad (30)$$

This finite-horizon capacity C_T is the operational capacity as given by Theorem 1 of [6] based on AEP and a random coding argument¹¹. Thus, we may focus only on the information rates in this paper and need not discuss coding in the operational sense.

¹⁰This \underline{v}^T is called innovations in [12], [36]; it should not be confused with the KF innovations in this paper.

¹¹One can also invoke Theorem 5.1 in [11] and the equivalence between directed information and mutual information in this case to claim that C_T is also the operation capacity.

To directly use the CP structure to construct a coding scheme is generally viewed as challenging for the following reasons. a) Its computation complexity grows faster than linearly with time ($O((T+1)^2)$ unknowns to be solved for each T), even though for each T the search of $\mathbf{K}_{\underline{v}}^{(T)}$ and \mathbf{B}_T can be posed as convex [19]. b) For each T the optimal $\mathbf{K}_{\underline{v}}^{(T)}$ and \mathbf{B}_T are not unique, (in fact there are an uncountable infinite number of optimizing solutions for each T , as can be easily seen from the $T = 1$ case); moreover, the optimal solution to coding length $(T+1)$ does not necessarily contain a part that is optimal to coding length T . Hence the search of optimal $\mathbf{K}_{\underline{v}}^{(T)}$ and \mathbf{B}_T for T is not likely to suggest what the optimal coding scheme could be for any other time horizon. c) In [6] the achievability of C_T is proven using a random coding argument, but a specific practical code has not been proposed or applied to the CP structure. Nevertheless, many insights can be obtained from the CP structure and it is also the starting point of our development.

B. The CP structure for the ISI Gaussian channel

In light of the correspondence relation between the colored Gaussian noise channel and the ISI channel \mathcal{F} , we can derive the CP coding structure for \mathcal{F} , which is obtained from (27) by introducing a new quantity \underline{z}^T as

$$\underline{z}^T := (\mathbf{I} + \mathbf{B}_T)^{-1} \underline{v}^T. \quad (31)$$

By $\underline{z}^T = \mathbf{Z}_T \underline{N}^T$ and $\tilde{y}^T = \mathbf{Z}_T \underline{y}^T$, we have

$$\begin{aligned} \underline{u}^T &= \mathbf{B}_T \mathbf{Z}_T \underline{N}^T + (\mathbf{I} + \mathbf{B}_T) \underline{z}^T \\ \underline{y}^T &= \mathbf{Z}_T^{-1} (\mathbf{I} + \mathbf{B}_T) \mathbf{Z}_T \underline{N}^T + \mathbf{Z}_T^{-1} (\mathbf{I} + \mathbf{B}_T) \underline{z}^T \\ &= \mathbf{Z}_T^{-1} (\mathbf{I} + \mathbf{B}_T) (\mathbf{Z}_T \underline{N}^T + \underline{z}^T). \end{aligned} \quad (32)$$

This implies that, the channel input \underline{u}^T can be represented as

$$\underline{u}^T = (\mathbf{I} + \mathbf{B}_T)^{-1} \mathbf{B}_T \mathbf{Z}_T \underline{y}^T + \underline{z}^T, \quad (33)$$

which leads to the block diagram in Fig. 4. Then the capacity C_T has the form:

$$\begin{aligned} C_T(\mathcal{P}) &= \sup \frac{1}{2(T+1)} \log \det \mathbf{K}_{\underline{y}}^{(T)} \\ &= \sup \frac{1}{2(T+1)} \log \det \left(\mathbf{Z}_T^{-1} (\mathbf{I} + \mathbf{B}_T) (\mathbf{Z}_T \mathbf{Z}_T' + \mathbf{K}_{\underline{z}}^{(T)}) (\mathbf{I} + \mathbf{B}_T)' \mathbf{Z}_T^{-1'} \right) \\ &= \sup \frac{1}{2(T+1)} \log \det (\mathbf{Z}_T \mathbf{Z}_T' + \mathbf{K}_{\underline{z}}^{(T)}) \end{aligned} \quad (34)$$

where the supremum is over the power constraint

$$P_T := \frac{1}{T+1} \text{tr}(\mathbf{B}_T \mathbf{Z}_T \mathbf{Z}_T' \mathbf{B}_T' + (\mathbf{I} + \mathbf{B}_T) \mathbf{K}_{\underline{z}}^{(T)} (\mathbf{I} + \mathbf{B}_T)') \leq \mathcal{P}. \quad (35)$$

The capacity in this form is equivalent to (29). Another form of the capacity based on the directed information, namely an input/output characterization, can be shown as equivalent to the above form; see Appendix II. One can also define the inverse function of $C_T(\mathcal{P})$ as $P_T(\mathcal{R})$, which is equal to the infimum power subject to a rate constraint

$$\frac{1}{2(T+1)} \log \det \mathbf{K}_{\underline{y}}^{(T)} \geq \mathcal{R}. \quad (36)$$

V. NECESSITY OF KF FOR OPTIMAL CODING

In this section, we consider a finite-horizon feedback coding structure over channel \mathcal{F} denoted $\mathbb{S} := \mathbb{S}(\mathcal{F})$, which is a variation of the CP structure. This variation is useful since: 1) searching over all possible parameters in the structure achieves C_T , that is, there is no loss of generality or optimality when focusing on this structure only; 2) we can show that to ensure power efficiency (to be explained), structure \mathbb{S} necessarily implements the KF algorithm. This implies that our KF characterization leads to a refinement to the CP structure.

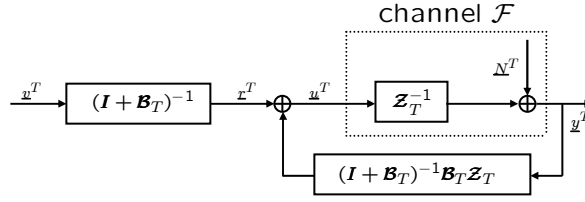


Fig. 4. The block diagram of the CP structure for ISI Gaussian channel \mathcal{F} .

A. Coding structure \mathbb{S}

Fig. 5 illustrates the coding structure \mathbb{S} , including the encoder and the *feedback generator*, which is a portion of the decoder. (How the decoder produces the estimate of the decoded message will be considered shortly.) Below, we fix the time horizon to span from time 0 to time T and describe \mathbb{S} .

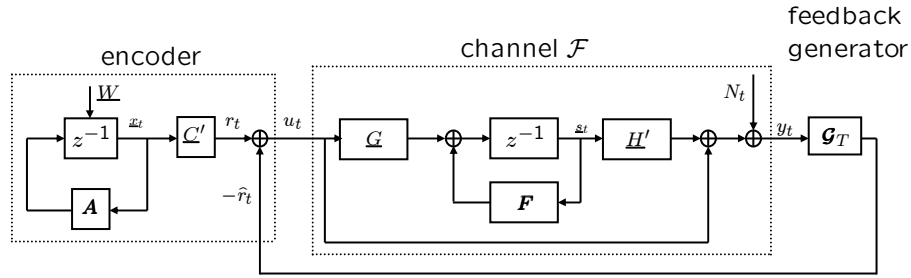


Fig. 5. Coding structure \mathbb{S} for channel \mathcal{F} .

Encoder: The encoder follows the dynamics

$$\text{Encoder: } \begin{cases} \underline{x}_{t+1} &= \mathbf{A} \underline{x}_t \\ r_t &= \underline{C}' \underline{x}_t \\ u_t &= r_t - \hat{r}_t. \end{cases} \quad (37)$$

where $\underline{x}_0 := \underline{W} \sim \mathcal{N}(\underline{0}, \mathbf{I}_{n+1})$. We assume that the encoder dimension $(n+1)$ is a fixed integer satisfying $0 \leq n \leq T$; $\mathbf{A} \in \mathbb{R}^{(n+1) \times (n+1)}$; $\underline{C}' \in \mathbb{R}^{n+1}$; and the assumption (A1) holds:

(A1): $(\mathbf{A}, \underline{C}')$ is observable.

We then let

$$\begin{aligned} \mathbf{\Gamma}_T(\mathbf{A}, \underline{C}') &:= \mathbf{\Gamma}_T := [\underline{C}', \mathbf{A}' \underline{C}', \dots, \mathbf{A}^{T'} \underline{C}']' \in \mathbb{R}^{(T+1) \times (n+1)} \\ \mathbf{K}_{\underline{r}}^{(T)}(\mathbf{A}, \underline{C}') &:= \mathbf{K}_{\underline{r}}^{(T)} := \mathbf{E}_{\underline{r}}^T \mathbf{\Gamma}_T^{-1} \in \mathbb{R}^{(T+1) \times (T+1)}. \end{aligned} \quad (38)$$

Therefore, $\mathbf{\Gamma}_n$ is the observability matrix for $(\mathbf{A}, \underline{C}')$ and is invertible, $\mathbf{\Gamma}_T$ has rank $(n+1)$, $\underline{r}^T = \mathbf{\Gamma}_T \underline{W}$, and $\mathbf{K}_{\underline{r}}^{(T)} = \mathbf{\Gamma}_T \mathbf{\Gamma}_T'$ with rank $(n+1)$.

Feedback generator: The feedback signal $(-\hat{r}_t)$ is generated through a feedback generator \mathcal{G}_T , i.e.

$$-\hat{\underline{r}}^T = \mathcal{G}_T \underline{y}^T, \quad (39)$$

where $\mathcal{G}_T \in \mathbb{R}^{(T+1) \times (T+1)}$ is a strictly lower triangular matrix, namely the output feedback is strictly causal.

Throughout the paper, the above assumptions on the encoder/decoder are always assumed if not otherwise specified. For future use, we compute the channel output as

$$\underline{y}^T = (\mathbf{I} - \mathbf{Z}_T^{-1} \mathcal{G}_T)^{-1} (\mathbf{Z}_T^{-1} \underline{r}^T + \underline{N}^T). \quad (40)$$

Definition 1. Consider the coding structure \mathbb{S} shown in Fig. 5. Define the constraint capacity

$$C_{T,n} := C_{T,n}(\mathcal{P}) := \sup_{\substack{\mathbf{A} \in \mathbb{R}^{(n+1) \times (n+1)}, \underline{\mathbf{C}}, \mathbf{G}_T, (A1) \\ \text{s.t. } \mathbf{E} \|\underline{\mathbf{u}}^T\|^2 / (T+1) \leq \mathcal{P}}} \frac{1}{T+1} I(\underline{\mathbf{W}}; \underline{\mathbf{y}}^T) \quad (41)$$

and define its inverse function as $P_{T,n}(\mathcal{R})$, that is,

$$P_{T,n} := P_{T,n}(\mathcal{R}) := \inf_{\substack{\mathbf{A} \in \mathbb{R}^{(n+1) \times (n+1)}, \underline{\mathbf{C}}, \mathbf{G}_T, (A1) \\ \text{s.t. } I(\underline{\mathbf{W}}; \underline{\mathbf{y}}^T) / (T+1) \geq \mathcal{R}}} \frac{1}{T+1} \mathbf{E} \|\underline{\mathbf{u}}^T\|^2 \quad (42)$$

In other words, $C_{T,n}$ is the finite-horizon information capacity for a *fixed encoder dimension* $(n+1)$, by searching over all admissible \mathbf{A} , $\underline{\mathbf{C}}$, and \mathbf{G}_T of appropriate dimensions. The pair $(\mathcal{P}, C_{T,n}(\mathcal{P}))$ and the pair $(P_{T,n}(\mathcal{R}), \mathcal{R})$ specify the *optimal tradeoff* between the channel input power and information rate for the communication problem with fixed encoder dimension.

B. Relation between the CP structure and the proposed structure \mathbb{S}

The coding structure \mathbb{S} over \mathcal{F} in Fig. 5 was motivated and is tightly associated with the CP structure over the ISI Gaussian channel \mathcal{F} in Fig. 4. Let $\underline{\mathbf{u}}^T(\mathbf{K}_r^{(T)}, \mathbf{B}_T)$ and $\underline{\mathbf{u}}^T(\mathbf{A}, \underline{\mathbf{C}}, \mathbf{G}_T)$ denote the input sequences generated by encoders with $(\mathbf{K}_r^{(T)}, \mathbf{B}_T)$ and $(\mathbf{A} \in \mathbb{R}^{(T+1) \times (T+1)}, \underline{\mathbf{C}}, \mathbf{G}_T)$, respectively.

Lemma 1. i) For any given pair $(\mathbf{K}_r^{(T)}, \mathbf{B}_T)$ with $\mathbf{K}_r^{(T)} > 0$, there exists an admissible triple $(\mathbf{A} \in \mathbb{R}^{(T+1) \times (T+1)}, \underline{\mathbf{C}}, \mathbf{G}_T)$ such that $\underline{\mathbf{u}}^T(\mathbf{K}_r^{(T)}, \mathbf{B}_T) = \underline{\mathbf{u}}^T(\mathbf{A}, \underline{\mathbf{C}}, \mathbf{G}_T)$; for any given pair $(\mathbf{K}_r^{(T)}, \mathbf{B}_T)$ with $\mathbf{K}_r^{(T)} \geq 0$ but $\mathbf{K}_r^{(T)} \not\geq 0$, there exists a sequence of admissible triples $\{(\mathbf{A}_i \in \mathbb{R}^{(T+1) \times (T+1)}, \underline{\mathbf{C}}_i, \mathbf{G}_{T,i})\}_{i=1}^\infty$ such that $\underline{\mathbf{u}}^T(\mathbf{K}_r^{(T)}, \mathbf{B}_T) = \lim_{i \rightarrow \infty} \underline{\mathbf{u}}^T(\mathbf{A}_i, \underline{\mathbf{C}}_i, \mathbf{G}_{T,i})$;

ii) For any given triple $(\mathbf{A} \in \mathbb{R}^{(T+1) \times (T+1)}, \underline{\mathbf{C}}, \mathbf{G}_T)$, there is an admissible pair $(\mathbf{K}_r^{(T)}, \mathbf{B}_T)$ such that $\underline{\mathbf{u}}^T(\mathbf{K}_r^{(T)}, \mathbf{B}_T) = \underline{\mathbf{u}}^T(\mathbf{A}, \underline{\mathbf{C}}, \mathbf{G}_T)$;

iii)

$$C_T(\mathcal{P}) = C_{T,T}(\mathcal{P}), \quad P_T(\mathcal{R}) = P_{T,T}(\mathcal{R}). \quad (43)$$

Proof: See Appendix III. \square

One advantage of considering the structure \mathbb{S} is that we can have the flexibility of allowing $T \geq n$, which makes it possible to increase the horizon length to infinity without increasing the dimension of \mathbf{A} , a useful step towards the KF characterization of the feedback communication problem.

In what follows, several refinements to the coding structure \mathbb{S} will be presented.

C. The presence of the KF

We first compute the mutual information in the aforementioned coding structure \mathbb{S} .

Proposition 1. Consider the structure \mathbb{S} in Fig. 5. Let $0 \leq n \leq T$, $(\mathbf{A}, \underline{\mathbf{C}})$ be observable with $\mathbf{A} \in \mathbb{R}^{(n+1) \times (n+1)}$ and \mathbf{G}_T be strictly lower triangular. Then

i) It holds that

$$\begin{aligned} I(\underline{\mathbf{W}}; \underline{\mathbf{y}}^T) &= I(\underline{\mathbf{x}}^T; \underline{\mathbf{y}}^T) \\ &= I(\underline{\mathbf{u}}^T \rightarrow \underline{\mathbf{y}}^T) \\ &= \frac{1}{2} \log \det \mathbf{K}_{\underline{\mathbf{y}}}^{(T)} \\ &= \frac{1}{2} \log \det (\mathbf{I} + \mathbf{Z}_T^{-1} \mathbf{K}_{\underline{\mathbf{x}}}^{(T)} \mathbf{Z}_T^{-1'}) \\ &= \frac{1}{2} \log \det (\mathbf{I} + \mathbf{Z}_T^{-1} \mathbf{\Gamma}_T \mathbf{\Gamma}_T' \mathbf{Z}_T^{-1'}); \end{aligned} \quad (44)$$

ii) $I(\underline{W}; \underline{y}^T)$ is independent of the feedback generator \mathcal{G}_T .

Proof: i)

$$\begin{aligned}
I(\underline{W}; \underline{y}^T) &= h(\underline{y}^T) - h(\underline{y}^T | \underline{W}) \\
&= h(\underline{y}^T) - h((\mathbf{I} - \mathcal{Z}_T^{-1} \mathcal{G}_T)^{-1} (\mathcal{Z}_T^{-1} \underline{x}^T + \underline{N}^T) | \underline{W}) \\
&\stackrel{(a)}{=} \frac{1}{2} \log \det(2\pi e \mathbf{K}_{\underline{y}}^{(T)}) - h(\underline{N}^T) \\
&\stackrel{(b)}{=} I(\underline{x}^T \rightarrow \underline{y}^T) \\
&= \frac{1}{2} \log \det \mathbf{K}_{\underline{y}}^{(T)} \\
&= \frac{1}{2} \log \det(\mathbf{I} + \mathcal{Z}_T^{-1} \mathbf{K}_{\underline{x}}^{(T)} \mathcal{Z}_T^{-1'}),
\end{aligned} \tag{45}$$

where (a) is due to $\underline{x}^T = \mathbf{\Gamma}_T \underline{W}$, $\det(\mathbf{AB}) = \det \mathbf{A} \det \mathbf{B}$, and $\det(\mathbf{I} - \mathcal{Z}_T^{-1} \mathcal{G}_T)^{-1} = 1$; and (b) follows from [14] or a direct computation of $I(\underline{x}^T \rightarrow \underline{y}^T)$. ii) It is clear from i) that $I(\underline{W}; \underline{y}^T)$ is independent of the feedback generator \mathcal{G}_T , and depends only on $\mathbf{K}_{\underline{x}}^{(T)}$, or equivalently on $(\mathbf{A}, \underline{C})$. \square

Remark 1. Though simple, Proposition 1 has interesting interpretations and implications. The first equality of i) shows that the mutual information between the message W and channel output \underline{y}^T is completely preserved in the mutual information between the message-carrying signal \underline{x}^T and channel output \underline{y}^T . The second equality shows that the directed information (cf. [11] and Appendix II) in this setup is equivalent to the message-output characterization based on the mutual information, which is convenient in many situations. The third equality involves the output covariance matrix, a link towards the Bode waterbed effect and the fundamental concept of the KF innovations (to be explored in subsequent sections). The rest of the proposition implies that, for the given channel \mathcal{Z}^{-1} , fixed $(\mathbf{A}, \underline{C})$ leads to a fixed information rate regardless of the feedback generator. In fact, the mutual information may be interpreted as anti-causal and independent of the strictly causal feedback generator. Hence the feedback generator \mathcal{G}_T has to be chosen to minimize the average channel input power in order to achieve the capacity (recalling that the capacity problem can be expressed as minimizing power while fixing the rate (42)), which necessitates a KF. Note that the infinite-horizon counterpart of this proposition was proven in [14].

Next we solve the optimal feedback generator for a fixed $(\mathbf{A}, \underline{C})$, which is essentially a KF. Denote the *optimal feedback generator* for a given $(\mathbf{A}, \underline{C})$ as $\mathcal{G}_T^*(\mathbf{A}, \underline{C})$, namely

$$\mathcal{G}_T^*(\mathbf{A}, \underline{C}) := \arg \inf_{\mathcal{G}_T} \frac{1}{T+1} \mathbf{E} \|\underline{u}^T(\mathbf{A}, \underline{C}, \mathcal{G}_T)\|^2. \tag{46}$$

By Proposition 1, we can define, for a fixed $(\mathbf{A}, \underline{C})$, the information rate across the channel to be

$$R_T(\mathbf{A}, \underline{C}) := \frac{I(\underline{W}; \underline{y}^T)}{T+1}. \tag{47}$$

Proposition 2. Consider coding structure \mathbb{S} in Fig. 5. Fix any $0 \leq n \leq T$. Then (recall the capacity definition $P_{T,n}(\mathcal{R})$ in (42))

i)

$$P_{T,n}(\mathcal{R}) = \inf_{\substack{\mathbf{A} \in \mathbb{R}^{(n+1) \times (n+1)}, \underline{C} \\ \text{s.t. } R_T(\mathbf{A}, \underline{C}) \geq \mathcal{R}}} \frac{1}{T+1} \mathbf{E} \|\underline{u}^T(\mathbf{A}, \underline{C}, \mathcal{G}_T^*(\mathbf{A}, \underline{C}))\|^2. \tag{48}$$

ii) The optimal feedback generator $\mathcal{G}_T^*(\mathbf{A}, \underline{C})$ is given by

$$\mathcal{G}_T^*(\mathbf{A}, \underline{C}) = -\hat{\mathcal{G}}_T^*(\mathbf{A}, \underline{C})(\mathbf{I} - \mathcal{Z}_T^{-1} \hat{\mathcal{G}}_T^*(\mathbf{A}, \underline{C}))^{-1}, \tag{49}$$

where $\hat{\mathcal{G}}_T^*(\mathbf{A}, \underline{C})$ is the one-step prediction MMSE estimator (Kalman filter) of \underline{x}^T given the noisy observation $\underline{y}^T := \mathcal{Z}_T^{-1} \underline{x}^T + \underline{N}^T$ (i.e. the optimal one-step prediction is $\hat{\underline{x}}^T = \hat{\mathcal{G}}_T^*(\mathbf{A}, \underline{C}) \underline{y}^T$), given by

$$\hat{\mathcal{G}}_T^*(\mathbf{A}, \underline{C}) := \arg \min_{\hat{\mathcal{G}}_T} \frac{1}{T+1} \mathbf{E} (\underline{x}^T - \hat{\mathcal{G}}_T \underline{y}^T)(\underline{x}^T - \hat{\mathcal{G}}_T \underline{y}^T)', \tag{50}$$

where $\hat{\mathbf{G}}_T$ is strictly lower triangular.

Fig. 6 (a) shows the associated estimation problem, (b) the KF $\hat{\mathbf{G}}_T^*(\mathbf{A}, \underline{\mathbf{C}})$ for (a), and (c) the state-space representation of the optimal feedback generator $\mathbf{G}_T^*(\mathbf{A}, \underline{\mathbf{C}})$ (see (61) and (65) for $\underline{L}_{1,t}$ and $\underline{L}_{2,t}$).

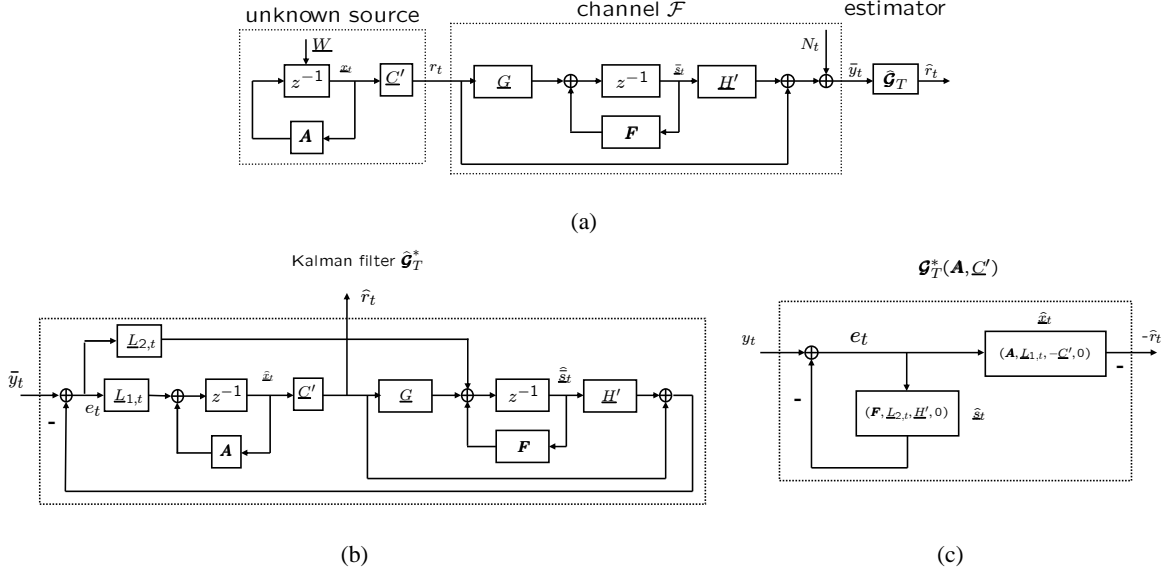


Fig. 6. (a) An estimation problem over channel \mathcal{F} . (b) The KF $\hat{\mathbf{G}}_T^*(\mathbf{A}, \underline{\mathbf{C}})$ for (a). (c) The KF-based feedback generator $\mathbf{G}_T^*(\mathbf{A}, \underline{\mathbf{C}})$ in state space. $(\mathbf{A}, \underline{L}_{1,t}, -\underline{C}', 0)$ with \hat{x}_t denotes a state-space representation with \hat{x}_t being its state at time t , and initial condition \hat{x}_0 being 0.

Remark 2. Proposition 2 reveals that, the minimization of channel input power in a feedback communication problem is equivalent to the minimization of MSE in an estimation problem. This equivalence yields a complete characterization (in terms of the KF algorithm) of the optimal feedback generator $\mathbf{G}_T^*(\mathbf{A}, \underline{\mathbf{C}})$ for any given $(\mathbf{A}, \underline{\mathbf{C}})$, as shown in Section VI-B. This proposition refines the CP structure as it shows that the CP structure necessarily contains a KF.

Remark 3. Proposition 2 i) implies that we may reformulate the problem of $C_{T,n}$ (or $P_{T,n}$) as a two-step problem: STEP 1: Fix $(\mathbf{A}, \underline{\mathbf{C}})$ (and hence fix the rate), and minimize the input power by searching over all possible feedback generator \mathbf{G} for the fixed $(\mathbf{A}, \underline{\mathbf{C}})$; STEP 2: Search over all possible $(\mathbf{A}, \underline{\mathbf{C}})$ subject to the rate constraint of $R_T(\mathbf{A}, \underline{\mathbf{C}}) \geq \mathcal{R}$. Thus, *one essential role of the feedback generator \mathbf{G}* for any fixed $(\mathbf{A}, \underline{\mathbf{C}})$ is to minimize the input power, which can be solved by considering the *equivalent* optimal estimation problem in Fig. 6 (a) whose solution is the KF. It also follows that $\mathbf{E}(u_t|y^{t-1}) = 0$, which implies no power waste due to a non-zero mean (cf. [12], Eq. (126)) and a center-of-gravity encoding rule (cf. [37]). The input generated by the KF-based feedback generator has the form

$$u_t = r_t - \mathbf{E}(r_t|y^{t-1}), \quad (51)$$

which is related to the optimal input distributions obtained by e.g. [12], [16], [38], [39].

We also remark that the necessity of the KF in the optimal coding scheme is not surprising, given various indications of the essential role of KF (or minimum mean squared-error estimators or MMSE estimators; or cheap control, its control theory equivalence; or the sum-product algorithm, its generalization) in optimal communication designs. See e.g. [12], [14], [33], [40]–[42]. The study of the KF in the feedback communication problem along the line of [42] may shed important insights on optimal communication problems and is under current investigation.

Proof: i) Notice that for any fixed $(\mathbf{A}, \underline{\mathbf{C}})$, $R_T(\mathbf{A}, \underline{\mathbf{C}})$ is fixed. Then from the definition of $P_{T,n}(\mathcal{R})$, we have

$$\begin{aligned} P_{T,n}(\mathcal{R}) &= \inf_{\substack{\mathbf{A}, \underline{\mathbf{C}}, \mathbf{G}_T \\ s.t. R_T(\mathbf{A}, \underline{\mathbf{C}}) \geq \mathcal{R}}} \frac{1}{T+1} \mathbf{E} \|\underline{\mathbf{u}}^T(\mathbf{A}, \underline{\mathbf{C}}, \mathbf{G}_T)\|^2 \\ &= \inf_{\substack{\mathbf{A}, \underline{\mathbf{C}}, (A1) \\ s.t. R_T(\mathbf{A}, \underline{\mathbf{C}}) \geq \mathcal{R}}} \inf_{(\mathbf{A}, \underline{\mathbf{C}}) \text{ fixed}, (A1), \mathbf{G}_T} \frac{1}{T+1} \mathbf{E} \|\underline{\mathbf{u}}^T(\mathbf{A}, \underline{\mathbf{C}}, \mathbf{G}_T)\|^2. \end{aligned} \quad (52)$$

Then i) follows from the definition of $\mathbf{G}_T^*(\mathbf{A}, \underline{\mathbf{C}})$.

ii) Note that for the coding structure \mathbb{S} , it holds that

$$\underline{\mathbf{u}}^T = \underline{\mathbf{r}}^T + (-\hat{\mathbf{r}}^T) = \underline{\mathbf{r}}^T + \mathbf{G}_T \underline{\mathbf{y}}^T. \quad (53)$$

Then, letting

$$\hat{\mathbf{G}}_T := -\mathbf{G}_T(\mathbf{I} - \mathbf{Z}_T^{-1}\mathbf{G}_T)^{-1} \quad (54)$$

and $\bar{\underline{\mathbf{y}}}^T := \mathbf{Z}_T^{-1}\underline{\mathbf{r}}^T + \underline{\mathbf{N}}^T$, we have $\mathbf{G}_T \underline{\mathbf{y}}^T = -\hat{\mathbf{G}}_T \bar{\underline{\mathbf{y}}}^T$. Therefore,

$$\begin{aligned} \mathbf{G}_T^*(\mathbf{A}, \underline{\mathbf{C}}) &= \arg \inf_{\mathbf{G}_T} \frac{1}{T+1} \mathbf{E}(\underline{\mathbf{r}}^T + \mathbf{G}_T \underline{\mathbf{y}}^T)(\underline{\mathbf{r}}^T + \mathbf{G}_T \underline{\mathbf{y}}^T)' \\ &= \arg \inf_{\hat{\mathbf{G}}_T} \frac{1}{T+1} \mathbf{E}(\underline{\mathbf{r}}^T - \hat{\mathbf{G}}_T \bar{\underline{\mathbf{y}}}^T)(\underline{\mathbf{r}}^T - \hat{\mathbf{G}}_T \bar{\underline{\mathbf{y}}}^T)'. \end{aligned} \quad (55)$$

The last equality implies that the optimal solution $\hat{\mathbf{G}}_T^*$ is the strictly causal MMSE estimator (with one-step prediction) of $\underline{\mathbf{r}}^T$ given $\bar{\underline{\mathbf{y}}}_T$; notice that $\hat{\mathbf{G}}_T$ is strictly lower triangular. It is well known that such an estimator can be implemented recursively in state-space as a KF (cf. [23], [43]). Finally, from the relation between \mathbf{G}_T and $\hat{\mathbf{G}}_T$, we obtain (49). The state-space representation of $\mathbf{G}_T^*(\mathbf{A}, \underline{\mathbf{C}})$, as illustrated in Fig. 6 (c), can be obtained from straightforward computation, as shown in Appendix I-A. \square

We remark that it is possible to derive a dynamic programming based solution ([11]) to compute $C_{T,n}$, and if we further employ the Markov property in [12] and the above KF-based characterization, we would reach a solution with complexity $O(T)$ for computing $C_{T,n}$ and C_T . However, we do not pursue along this line in this paper as it is beyond the main scope of this paper.

VI. CONNECTIONS AMONG FEEDBACK COMMUNICATION, ESTIMATION, AND FEEDBACK CONTROL

We have shown that in the coding structure \mathbb{S} , to ensure power efficiency for a fixed $(\mathbf{A}, \underline{\mathbf{C}})$, one needs to design a KF-based feedback generator. The KF immediately links the feedback communication problem to estimation and control problems. In this section, we present a *unified representation* of the *optimal* coding structure \mathbb{S}^* (i.e., \mathbb{S}^* is \mathbb{S} but with \mathbf{G} being chosen as $\mathbf{G}^*(\mathbf{A}, \underline{\mathbf{C}})$), its estimation theory counterpart, and its control theory counterpart. Then in the next section we will establish relation among the information theory quantities, estimation theory quantities, and control theory quantities.

A. Unified representation of feedback coding system, KF, and cheap control

Coding structure \mathbb{S}^*

The optimal feedback generator for a given $(\mathbf{A}, \underline{\mathbf{C}})$ is solved in (49), see Fig. 6 (c) for its structure. We can then obtain a state-space representation of the optimal feedback generator $\mathcal{G}_T^*(\mathbf{A}, \underline{\mathbf{C}})$, and describe the coding structure \mathbb{S}^* which contains $\mathcal{G}_T^*(\mathbf{A}, \underline{\mathbf{C}})$ as

$$\text{coding structure } \mathbb{S}^* \left\{ \begin{array}{ll} \left. \begin{array}{l} \underline{x}_{t+1} = \mathbf{A}\underline{x}_t \\ r_t = \underline{\mathbf{C}}' \underline{x}_t \\ u_t = r_t - \hat{r}_t \end{array} \right\} & \text{encoder} \\ \left. \begin{array}{l} \underline{s}_{t+1} = \mathbf{F}\underline{s}_t + \underline{\mathbf{G}}u_t \\ y_t = \underline{\mathbf{H}}' \underline{s}_t + u_t + N_t \end{array} \right\} & \text{channel } \mathcal{F} \\ \left. \begin{array}{l} \hat{\underline{s}}_{t+1} = \mathbf{F}\hat{\underline{s}}_t + \underline{\mathbf{L}}_{2,t}e_t \\ e_t = y_t - \underline{\mathbf{H}}' \hat{\underline{s}}_t \\ \hat{\underline{x}}_{t+1} = \mathbf{A}\hat{\underline{x}}_t + \underline{\mathbf{L}}_{1,t}e_t \\ -\hat{r}_t = -\underline{\mathbf{C}}' \hat{\underline{x}}_t \end{array} \right\} & \text{optimal feedback generator } \mathcal{G}^*(\mathbf{A}, \underline{\mathbf{C}}) \end{array} \right. \quad (56)$$

with $\underline{x}_0 = \underline{W}$ unknown, $s_0 = \hat{s}_0 = \underline{0}$, and $\hat{x}_0 = \underline{0}$. Here $\underline{\mathbf{L}}_{1,t} \in \mathbb{R}^{n+1}$ and $\underline{\mathbf{L}}_{2,t} \in \mathbb{R}^m$ are the time-varying KF gains specified in (64). See Appendix I for the derivation of a state-space representation of $\mathcal{G}_T^*(\mathbf{A}, \underline{\mathbf{C}})$.

The estimation system

The estimation system in Fig. 6 (a) and (b) consists of three parts: the unknown source \underline{x}^T to be estimated or tracked, the channel \mathcal{F} (without output feedback), and the estimator, which we choose as the KF $\hat{\mathcal{G}}^*$; we assume that $(\mathbf{A}, \underline{\mathbf{C}})$ is fixed and known to the estimator and hence the randomness in \underline{x}^T comes from the initial condition of \underline{x}^T . The system is described in state-space as

$$\text{estimation system: } \left\{ \begin{array}{ll} \left. \begin{array}{l} \underline{x}_{t+1} = \mathbf{A}\underline{x}_t \\ r_t = \underline{\mathbf{C}}' \underline{x}_t \end{array} \right\} & \text{unknown source} \\ \left. \begin{array}{l} \bar{\underline{s}}_{t+1} = \mathbf{F}\bar{\underline{s}}_t + \underline{\mathbf{G}}r_t \\ \bar{y}_t = \underline{\mathbf{H}}' \bar{\underline{s}}_t + r_t + N_t \end{array} \right\} & \text{channel } \mathcal{F} \\ \left. \begin{array}{l} \hat{\underline{x}}_{t+1} = \mathbf{A}\hat{\underline{x}}_t + \underline{\mathbf{L}}_{1,t}e_t \\ \hat{r}_t = \underline{\mathbf{C}}' \hat{\underline{x}}_t \\ \hat{\underline{s}}_{t+1} = \mathbf{F}\hat{\underline{s}}_t + \underline{\mathbf{G}}\hat{r}_t + \underline{\mathbf{L}}_{2,t}e_t \\ e_t = \bar{y}_t - \underline{\mathbf{H}}' \hat{\underline{s}}_t - \hat{r}_t \end{array} \right\} & \text{Kalman filter } \hat{\mathcal{G}}^*(\mathbf{A}, \underline{\mathbf{C}}) \end{array} \right. \quad (57)$$

with $\underline{x}_0 = \underline{W}$, $\bar{s}_0 = \hat{s}_0 = \underline{0}$, and $\hat{x}_0 = \underline{0}$. To write this in a more compact form, define

$$\begin{aligned} \bar{\underline{\mathbf{X}}}_t &:= \begin{bmatrix} \underline{x}_t \\ \bar{\underline{s}}_t \end{bmatrix} \\ \hat{\underline{\mathbf{X}}}_t &:= \begin{bmatrix} \hat{\underline{x}}_t \\ \hat{\underline{s}}_t \end{bmatrix} \\ \mathbf{A} &:= \left[\begin{array}{c|c} \mathbf{A} & \mathbf{0} \\ \hline \underline{\mathbf{G}}\underline{\mathbf{C}}' & \mathbf{F} \end{array} \right] \\ \underline{\mathbf{C}} &:= \begin{bmatrix} \underline{\mathbf{C}} \\ \underline{\mathbf{H}} \end{bmatrix} \\ \underline{\mathbf{L}}_t &:= \begin{bmatrix} \underline{\mathbf{L}}_{1,t} \\ \underline{\mathbf{L}}_{2,t} \end{bmatrix}. \end{aligned} \quad (58)$$

Then we have

$$\text{estimation system: } \left\{ \begin{array}{ll} \left. \begin{array}{l} \bar{\underline{\mathbf{X}}}_{t+1} = \mathbf{A}\bar{\underline{\mathbf{X}}}_t \\ \bar{y}_t = \underline{\mathbf{C}}'\bar{\underline{\mathbf{X}}}_t + N_t \end{array} \right\} & \text{unknown source and channel } \mathcal{F} \\ \left. \begin{array}{l} \hat{\underline{\mathbf{X}}}_{t+1} = \mathbf{A}\hat{\underline{\mathbf{X}}}_t + \underline{\mathbf{L}}_t e_t \\ e_t = \bar{y}_t - \underline{\mathbf{C}}'\hat{\underline{\mathbf{X}}}_t \end{array} \right\} & \text{Kalman filter } \hat{\mathcal{G}}^*(\mathbf{A}, \underline{\mathbf{C}}) \end{array} \right. \quad (59)$$

with $\underline{\underline{X}}_0 = [\underline{W}', \underline{0}']'$ and $\hat{\underline{\underline{X}}}_0 = \underline{0}$.

It can be easily shown that r_t , \hat{r}_t , e_t , \underline{x}_t , and $\hat{\underline{x}}_t$ in (57) and (56) are equal, respectively, and it holds that for any t ,

$$\underline{s}_t - \hat{\underline{s}}_t = \bar{\underline{s}}_t - \hat{\bar{\underline{s}}}_t, \quad (60)$$

which leads to the following unified representation as a control system.

The unified representation: A cheap control problem

Define

$$\begin{aligned} \tilde{\underline{x}}_t &:= \underline{x}_t - \hat{\underline{x}}_t \\ \tilde{\underline{s}}_t &:= \underline{s}_t - \hat{\underline{s}}_t = \bar{\underline{s}}_t - \hat{\bar{\underline{s}}}_t \\ \underline{\underline{X}}_t &:= \begin{bmatrix} \tilde{\underline{x}}_t \\ \tilde{\underline{s}}_t \end{bmatrix} = \underline{\underline{X}}_t - \hat{\underline{\underline{X}}}_t \\ \underline{\underline{X}}_0 &:= \begin{bmatrix} \underline{W}' \\ \underline{0} \end{bmatrix} \\ \underline{\underline{D}} &:= \begin{bmatrix} \underline{C}' \\ \underline{0} \end{bmatrix}. \end{aligned} \quad (61)$$

Note that $\underline{\underline{X}}_t$ is the estimation error for $\hat{\underline{\underline{X}}}_t$. Substituting (61) to (57) and (56), we obtain that both systems become

$$\text{control system: } \begin{cases} \underline{\underline{X}}_{t+1} = (\underline{\mathbf{A}} - \underline{L}_t \underline{\mathbf{C}}') \underline{\underline{X}}_t - \underline{L}_t N_t = \underline{\mathbf{A}} \underline{\underline{X}}_t - \underline{L}_t e_t & \text{(state evolution)} \\ e_t = \underline{\mathbf{C}}' \underline{\underline{X}}_t + N_t & \text{(noisy measurement)} \\ u_t = \underline{\mathbf{D}}' \underline{\underline{X}}_t & \text{(regulated output)} \end{cases} \quad (62)$$

See Fig. 7 for the block diagrams. It is a control system where we want to minimize the power of the regulated output u by appropriately choosing \underline{L}_t . More specifically, one may view e_t as the noisy measurement which is also the input to the controller, $(-\underline{L}_t)$ as the time-varying controller gain, $(-\underline{L}_t e_t)$ as the controller's output which is also the input to the system with state $\underline{\underline{X}}_t$. The objective is to minimize $\mathbf{E} \|\underline{u}^T\|^2$; more formally we want to solve

$$P_{T,n}(\underline{\mathbf{A}}, \underline{\mathbf{C}}) := \min_{\substack{\underline{L}_0, \dots, \underline{L}_T \\ \text{s.t. (62)}}} \frac{1}{T+1} \mathbf{E} \|\underline{u}^T\|^2 \quad (63)$$

in which $\underline{\mathbf{A}}, \underline{\mathbf{C}}$ are given and \underline{W} unknown. Note that the control effort is “free” as there is no direct penalty on the controller's output $(-\underline{L}_t e_t)$. This is a *cheap control* problem, which is useful for us to characterize the steady-state solution and it is equivalent to the KF problem (see [44]; also see [21] for the discussion of cheap control and the closely related expensive control and minimum-energy control).

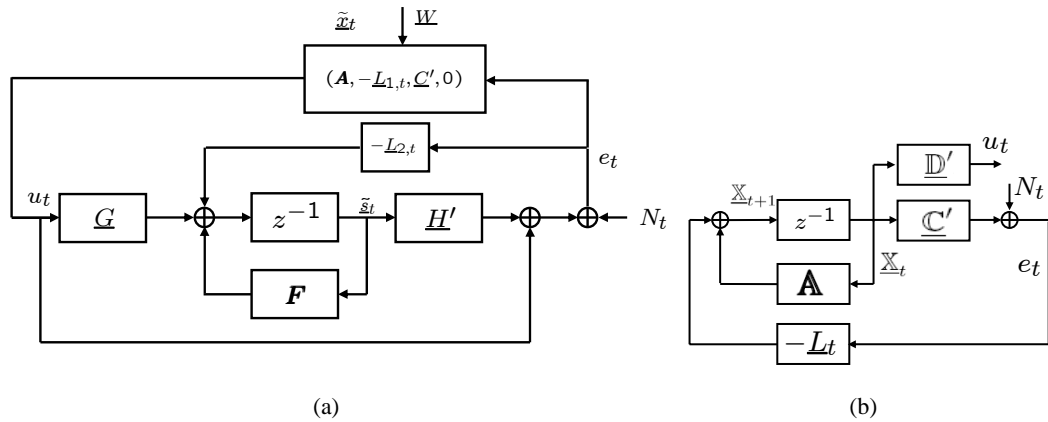


Fig. 7. Two equivalent block diagrams for the cheap control system. In (a) the block $(\underline{\mathbf{A}}, -\underline{L}_{1,t}, \underline{\mathbf{C}}', 0)$ denotes the state-space representation with $\tilde{\underline{x}}_t$ and \underline{W} being its states at time t and at time 0, respectively.

The signal e_t in (62) is the *KF innovation* or simply *innovation*¹². One fact is that $\{e_t\}$ is a white process, that is, its covariance matrix $\mathbf{K}_e^{(T)}$ is a diagonal matrix. Another fact is that \underline{e}^T and \underline{y}^T determine each other causally, and we can easily verify that $h(\underline{e}^T) = h(\underline{y}^T)$ and $\det \mathbf{K}_{\underline{y}}^{(T)} = \det \mathbf{K}_{\underline{e}}^{(T)}$. We remark that (62) is the *innovations representation* of the KF (cf. [23]).

For each t , the optimal \underline{L}_t is determined as

$$\underline{L}_t := \begin{bmatrix} \underline{L}_{1,t} \\ \underline{L}_{2,t} \end{bmatrix} := \frac{\mathbf{A}\Sigma_t\mathbf{C}}{K_{e,t}}, \quad (64)$$

where $\Sigma_t := \mathbf{E}\underline{\mathbf{X}}_t\underline{\mathbf{X}}_t'$, $K_{e,t} := \mathbf{E}(e_t)^2 = \mathbf{C}'\Sigma_t\mathbf{C} + 1$, and the error covariance matrix Σ_t satisfies the Riccati recursion

$$\Sigma_{t+1} = \mathbf{A}\Sigma_t\mathbf{A}' - \frac{\mathbf{A}\Sigma_t\mathbf{C}\mathbf{C}'\Sigma_t\mathbf{A}'}{\mathbf{C}'\Sigma_t\mathbf{C} + 1} \quad (65)$$

with initial condition

$$\Sigma_0 := \begin{bmatrix} \mathbf{I}_{n+1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad (66)$$

This completes the description of the optimal feedback generator for a given (\mathbf{A}, \mathbf{C}) .

The existence of one unified expression for three different systems (57), (56), and (62) is because the first two are actually two different non-minimal realizations of the third. The input-output mappings from \underline{N}^T to \underline{e}^T in the three systems are T -equivalent (see Appendix I-B). Thus we say that the three problems, the optimal estimation problem, the optimal feedback generator problem, and the cheap control problem, are *equivalent* in the sense that, if any one of the problems is solved, then the other two are solved. Since the estimation problem and the control problem are well studied, the equivalence can sometimes facilitate our study of the communication problem. Particularly, the formulation (62) yields alternative expressions for the mutual information and average channel input power in the feedback communication problem, as we see in the next section.

We further illustrate the relation of the estimation system and the communication system in Fig. 8, in which (b) is obtained from (a) by subtracting \hat{r}_t from the channel input and adding $(\mathbf{Z}_T^{-1}\hat{r}_t)$ back to the channel output, which does not affect the input, state, and output of $\hat{\mathbf{G}}_T^*$. It is clearly seen from the block diagram manipulations that *the minimization of channel input power in feedback communication problem becomes the minimization of MSE in the estimation problem*. This generalizes the observation we made regarding how to obtain a coding structure from a KF over an AWGN channel (as shown in Fig. 1) to more general Gaussian channels.

B. Roles of the KF algorithm in feedback communication

We have seen that the KF algorithm is necessary to ensure the power efficiency in feedback communication. Here we show that it is also needed to recover the transmitted signal $\underline{x}_0 := \underline{W}$.

The estimation of \underline{x}_0 is a (an anti-causal) smoothing problem; more specifically, a fixed-point smoothing problem (cf. e.g. Ch. 10 of [23]), whose solution is typically easily obtained by studying the innovations process of the KF used for prediction. Note that $\underline{\mathbf{X}}_0 := [\underline{x}_0', \underline{z}_0'] := [\underline{W}', \underline{0}']'$, and hence the smoothed estimate for \underline{x}_0 can be obtained by the smoothed estimate of $\underline{\mathbf{X}}_0$ (the constraint that $\underline{z}_0 := \underline{0}$ should be automatically satisfied in the smoothing problem solution). Denote $\hat{\underline{\mathbf{X}}}_{0|t} := \mathbf{E}(\underline{\mathbf{X}}_0|\underline{y}^t)$ and $\hat{\underline{x}}_{0|t} := \mathbf{E}(\underline{W}|\underline{y}^t)$. The solution is given below. Denote the closed-loop state transition matrices as $\Phi(t) := \mathbf{A}_{cl}(t-1)\mathbf{A}_{cl}(t-2)\cdots\mathbf{A}_{cl}(0)$ if $t > 0$ and $\Phi(0) := \mathbf{I}$, where $\mathbf{A}_{cl}(t) := \mathbf{A} - \underline{L}_t\mathbf{C}'$, and $\phi(t) := \mathbf{A}_{cl}(t-1)\mathbf{A}_{cl}(t-2)\cdots\mathbf{A}_{cl}(0)$ if $t > 0$ and $\phi(0) := \mathbf{I}$, where $\mathbf{A}_{cl}(t) := \mathbf{A} - \underline{L}_{1,t}\mathbf{C}'$. (It holds that $\phi(t)$ is the upper left block of $\Phi(t)$.) Then the smoothing equations are (see Problem 10.1 in [23])

$$\begin{aligned} \hat{\underline{\mathbf{X}}}_{0,t} &= \hat{\underline{\mathbf{X}}}_{0,t-1} + \Sigma_0\Phi'(t)\mathbf{C}K_{e,t}^{-1}e_t \\ \hat{\underline{x}}_{0,t} &= \hat{\underline{x}}_{0,t-1} + \phi'(t)\mathbf{C}K_{e,t}^{-1}e_t, \end{aligned} \quad (67)$$

¹²The innovation defined here is consistent with the Kalman filtering literature but different from that defined in [6] or [12].

by Mitter and Newton in [42]. It is also interesting to note that, though for different classes of channels, different optimal coding schemes have been derived along different directions, these schemes can be universally interpreted in terms of KF of appropriate forms; see [22]. Thus, we consider that the KF acts as a “unifier” for feedback communication schemes over various channels.

Finally, our study on the coding structure \mathbb{S} also refines the CP structure. Indeed, we conclude that the CP structure needs to have a KF inside. We may further determine the optimal form of \mathbf{B}_T . From (120) and (49), we have that

$$\mathbf{B}_T^* = -\hat{\mathbf{G}}_T^*(\mathbf{A}, \underline{\mathbf{C}})\mathbf{Z}_T^{-1}, \quad (71)$$

where $\hat{\mathbf{G}}_T^*(\mathbf{A}, \underline{\mathbf{C}})$ is the KF given in (57). Therefore, to achieve C_T in the CP structure, it is sufficient to search $(\mathbf{K}_v^{(T)}, \mathbf{B}_T)$ in the form of

$$\begin{aligned} \mathbf{K}_v^{(T)} &:= (\mathbf{I} - \hat{\mathbf{G}}_T^*(\mathbf{A}, \underline{\mathbf{C}})\mathbf{Z}_T^{-1})\mathbf{\Gamma}_T(\mathbf{A}, \underline{\mathbf{C}})\mathbf{\Gamma}_T(\mathbf{A}, \underline{\mathbf{C}})'(\mathbf{I} - \hat{\mathbf{G}}_T^*(\mathbf{A}, \underline{\mathbf{C}})\mathbf{Z}_T^{-1})' \\ \mathbf{B}_T^* &:= -\hat{\mathbf{G}}_T^*(\mathbf{A}, \underline{\mathbf{C}})\mathbf{Z}_T^{-1}. \end{aligned} \quad (72)$$

VII. CONNECTIONS OF FUNDAMENTAL LIMITATIONS

In this section, we discuss the connections of fundamental limitations. These limitations involve the mutual information in the feedback communication system, the Fisher information, MMSE, and CRB in the estimation system, and the Bode sensitivity integral in the feedback control system. We show that one limitation may be expressed in terms of the others, as a consequence of the equivalence established above.

A. Fisher information matrix (FIM), CRB, and Bode-type sensitivity integral (sum)

Let us first recall the general definitions of MMSE, Fisher information matrix (FIM), and CRB:

$$\text{MMSE}_{\underline{W}} := \mathbf{E}(\underline{W} - \hat{\underline{W}})(\underline{W} - \hat{\underline{W}})', \quad (73)$$

where $\hat{\underline{W}} := \mathbf{E}(\underline{W}|\underline{y})$ is the MMSE estimator of \underline{W} based on noisy observation \underline{y} ;

$$\begin{aligned} \mathcal{I}_{\underline{W}} &:= \mathbf{E} \left(\frac{\partial \log p_{\underline{W}, \underline{y}}(\underline{W}, \underline{y})}{\partial \underline{W}} \right)^2 \\ &= -\mathbf{E} \left(\frac{\partial^2 \log p_{\underline{W}, \underline{y}}(\underline{W}, \underline{y})}{\partial \underline{W}^2} \right) \end{aligned} \quad (74)$$

to be the (Bayesian) FIM, where $p_{\underline{W}, \underline{y}}(\underline{W}, \underline{y})$ is the joint density of \underline{W} and \underline{y} ; and

$$\text{CRB}_{\underline{W}} := \mathcal{I}_{\underline{W}}^{-1} \quad (75)$$

to be the (Bayesian) CRB [24]. Note that it always holds, as a *fundamental limitation in estimation theory*, that

$$\text{MSE}_{\underline{W}} \geq \text{CRB}_{\underline{W}}, \quad (76)$$

regardless of how one designs the estimator [24]. This inequality is referred to as the information inequality, Cramer-Rao inequality, or van Trees inequality¹⁴.

The Bode sensitivity integral is a *fundamental limitation in feedback control* (typically in steady state). Simply put, for any feedback design, the sensitivity of the output to exogenous disturbance cannot be made small uniformly over all frequencies since the sensitivity transfer function’s power spectrum in log scale sums up (integrates) to be constant. See Section IX-B and [29]. A similar limitation holds in finite horizon as we now show.

¹⁴Some authors distinguish the Cramer-Rao inequality and van Trees inequality by restricting the former to be non-Bayesian and unbiased and the latter to be Bayesian and possibly biased.

As

$$\underline{y}^T = (\mathbf{I} - \underline{\mathbf{Z}}_T^{-1} \underline{\mathbf{G}}_T)^{-1} (\underline{\mathbf{Z}}_T^{-1} \underline{\mathbf{r}}^T + \underline{\mathbf{N}}^T), \quad (77)$$

the sensitivity of channel output \underline{y}^T to noise \underline{N}^T is $\mathcal{S}_T := (\mathbf{I} - \underline{\mathbf{Z}}_T^{-1} \underline{\mathbf{G}}_T)^{-1}$. It is then easily seen that, if the spectrum of $\mathcal{S}_T \mathcal{S}_T'$ is $\{\lambda_i\}_{i=1}^{T+1}$, then

$$\sum_{i=1}^{T+1} \log \lambda_i = 0, \quad (78)$$

which holds valid regardless of the choice of feedback generator $\underline{\mathbf{G}}_T$, including the case that there is no feedback (i.e. open loop). Thus, the effect of noise \underline{N}^T cannot be made arbitrarily small in the measurements \underline{y}^T , which may be viewed as a fundamental limitation of noise (or disturbance) suppression.

Since the noise \underline{N}^T is normalized, one may also define the sensitivity based on the spectrum of $\mathbf{K}_{\underline{y}}^{(T)}$ or on the innovation process variance $K_{e,t}$. Let

$$\text{BI}_T := \frac{1}{2} \sum_{i=1}^{T+1} \log \lambda_i(\mathbf{K}_{\underline{y}}^{(T)}) = \sum_{t=0}^T \log K_{e,t}, \quad (79)$$

which is easily seen independent of any causal feedback and is the finite-horizon counterpart of the widely known Bode sensitivity integral of infinite-horizon.

B. Expressions for mutual information and channel input power

We have the following proposition.

Proposition 3. Consider the coding structure \mathbb{S} . For any fixed $0 \leq n \leq T$ and observable $(\mathbf{A}, \underline{\mathbf{C}}')$ with $\mathbf{A} \in \mathbb{R}^{(n+1) \times (n+1)}$, it holds that

i)

$$\begin{aligned} I(\underline{W}; \underline{y}^T) &= \text{BI}_T = \frac{1}{2} \sum_{t=0}^T \log K_{e,t} \\ &= \frac{1}{2} \sum_{t=0}^T \log (\underline{\mathbf{C}}' \underline{\Sigma}_t \underline{\mathbf{C}} + 1) \\ &= \frac{1}{2} \log \det \text{MMSE}_{\underline{W}, T}^{-1} \\ &= \frac{1}{2} \log \det \mathbf{I}_{\underline{W}, T} \\ &= \frac{1}{2} \log \det \text{CRB}_{\underline{W}, T}^{-1}; \end{aligned} \quad (80)$$

ii)

$$\begin{aligned} P_{T,n}(\mathbf{A}, \underline{\mathbf{C}}) &= \frac{1}{T+1} \sum_{t=0}^T \underline{\mathbf{D}}' \underline{\Sigma}_t \underline{\mathbf{D}} \\ &= \frac{1}{T+1} \text{trace}(\text{PMMSE}_{\underline{\mathbf{r}}, T}) \\ &= \frac{1}{T+1} \sum_{t=0}^T \underline{\mathbf{C}}' \mathbf{A}^t \text{MMSE}_{\underline{W}, t} \mathbf{A}^{t'} \underline{\mathbf{C}}, \end{aligned} \quad (81)$$

where $\text{MMSE}_{\underline{W}, T}$ is the minimum MSE of \underline{W} at time T , $\text{CMMSE}_{\underline{\mathbf{r}}, T}$ is the causal minimum MSE of $\underline{\mathbf{r}}^T$ at time T , $\mathbf{I}_{\underline{W}, T}$ is the Bayesian Fisher information matrix of \underline{W} at time T for the estimation system (57), and $\text{CRB}_{\underline{W}, T}$ is the Bayesian CRB of \underline{W} at time T .

Note that $\text{PMMSE}_{\underline{\mathbf{r}}, T} := \mathbf{E}(\hat{\underline{\mathbf{r}}}^T - \underline{\mathbf{r}}^T)(\hat{\underline{\mathbf{r}}}^T - \underline{\mathbf{r}}^T)'$, in which $\hat{\underline{\mathbf{r}}}^T = [\hat{r}_0, \dots, \hat{r}_T]$ contains the (strictly causal) estimates with one-step prediction $\hat{r}_t := \underline{\mathbf{C}}' \mathbf{E}(\underline{\mathbf{x}}_t | \underline{y}^{t-1})$ for $t = 0, \dots, T$.

Remark 5. This proposition connects the mutual information to the Bode sensitivity integral of the associated control problem and to the innovations process, Fisher information, (minimum) MSE, and CRB of the associated estimation problem. Note that any mutual information larger than the value given above is not possible regardless of how one designs the feedback generator, and how much mutual information we may obtain is limited by the control problem fundamental limitations and by how well the estimation can be done and hence by the Fisher information, MMSE, and CRB. Thus the fundamental limitation in feedback communication is linked to the fundamental limitations in control and estimation.

This proposition also shows that the spectrum of the output covariance matrix or the innovation variance cannot be made large or small uniformly, which may be viewed as the finite-horizon, time domain counterpart of the Bode sensitivity integral in the steady state and frequency domain. Notice that so far the estimation problem and control problem do not rely on asymptotic notions such as stability (stability was used to establish the Bode-Shannon connections between feedback communication and feedback stabilization in steady state [14]).

As a side note, if one defines the complementary sensitivity as $\mathcal{T}_T := \mathbf{Z}_T^{-1} \mathbf{G}_T (\mathbf{I} - \mathbf{Z}_T^{-1} \mathbf{G}_T)^{-1}$, it still holds that $\mathcal{S}_T - \mathcal{T}_T = \mathbf{I}$, which resembles the fundamental algebraic tradeoff in the steady state and frequency domain (cf. [29]).

Proof: i) First we simply notice that $h(\underline{y}^T) = h(\underline{e}^T)$, and $K_{e,t} = \mathbb{C}' \Sigma_t \mathbb{C} + 1$. Next, to find MMSE of \underline{W} , note that in Fig. 6 (a)

$$\underline{y}^T = \mathbf{Z}_T^{-1} \Gamma_T \underline{W} + \underline{N}^T \quad (82)$$

and that $\underline{W} \sim \mathcal{N}(\underline{0}, \mathbf{I})$, $\underline{N}^T \sim \mathcal{N}(\underline{0}, \mathbf{I})$. Thus, by [43] we have

$$\text{MMSE}_{\underline{W},t} = (\mathbf{I} + \Gamma_T' \mathbf{Z}_T^{-1'} \mathbf{Z}_T^{-1} \Gamma_T)^{-1} = \mathbf{I} - \Gamma_T' (\mathbf{Z}_T \mathbf{Z}_T' + \Gamma_T' \Gamma_T)^{-1} \Gamma_T, \quad (83)$$

yielding

$$\begin{aligned} \det \text{MMSE}_{\underline{W},t} &= \det(\mathbf{I} + \mathbf{Z}_T^{-1} \Gamma_T \Gamma_T' \mathbf{Z}_T^{-1'})^{-1} \\ &= \det(\mathbf{I} + \mathbf{Z}_T^{-1} \mathbf{K}_r^{(T)} \mathbf{Z}_T^{-1'})^{-1} = \det(\mathbf{K}_r^{(T)} + \mathbf{K}_{\underline{Z}}^{(T)})^{-1}. \end{aligned} \quad (84)$$

Besides, from Section 2.4 in [24] we can directly compute the FIM of \underline{W} to be $(\mathbf{I} + \Gamma_T' \mathbf{Z}_T^{-1'} \mathbf{Z}_T^{-1} \Gamma_T)$. Then i) follows from Proposition 1 and (62).

ii) Since $u_t = \mathbb{D}' \underline{\mathbf{x}}_t = \mathbb{C}' \hat{\underline{\mathbf{x}}}_t = r_t - \hat{r}_t$ and $\mathbf{E} \hat{\underline{\mathbf{x}}}_t \hat{\underline{\mathbf{x}}}_t' = \mathbf{A}^t \text{MMSE}_{\underline{W},t} \mathbf{A}^{t'}$, we have $\mathbf{E}(u_t)^2 = \mathbb{D}' \Sigma_t \mathbb{D} = \mathbb{C}' \mathbf{E} \hat{\underline{\mathbf{x}}}_t \hat{\underline{\mathbf{x}}}_t' \mathbb{C} = \mathbf{E}(r_t - \hat{r}_t)^2$, and then ii) follows. \square

C. Connections of the fundamental tradeoffs

The above fundamental limitations are based on *one fixed* (\mathbf{A}, \mathbb{C}) with $\mathbf{A} \in \mathbb{R}^{n \times n}$. Searching over *all* admissible (\mathbf{A}, \mathbb{C}) with $\mathbf{A} \in \mathbb{R}^{n \times n}$ for all $n \leq T$, one can obtain the optimal tradeoffs for feedback communication, estimation, and feedback control, as well as the corresponding relation among these tradeoffs. Note that the linear scheme with (\mathbf{A}, \mathbb{C}) can attain the optimal tradeoffs as we have established in the feedback communication system (see Proposition 2), and hence the optimal tradeoffs obtained by searching over all admissible (\mathbf{A}, \mathbb{C}) are indeed the optimal tradeoffs over all (possibly nonlinear, provided relevant quantities are well defined) feedback communication designs, estimator designs, and feedback control designs. These fundamental tradeoffs are elaborated below.

The fundamental tradeoff in the feedback communication problem over the channel \mathcal{F} for finite-horizon from time 0 to time T is the capacity $C_{T,T}(\mathcal{P})$ (or $P_{T,T}(\mathcal{R})$, see Definition 1) in the form of the optimal power-rate pair. (As indicated by Proposition 2, searching over all admissible (\mathbf{A}, \mathbb{C}) achieves the capacity.) That is, we have:

(T1) Optimal Feedback Communication Tradeoff: Given the channel \mathcal{F} with one-step delayed output feedback and an average channel input power \mathcal{P} , the achievable information rate $R_T(\mathbf{f}, \mathcal{P})$ cannot be higher than a constant $C_{T,T}(\mathcal{P})$ for any feedback communication design \mathbf{f} ; here $R_T(\mathbf{f}, \mathcal{P}) := \frac{1}{T+1} I(\underline{u}^T(\mathbf{f}) \rightarrow \underline{y}^T(\mathbf{f}))$ is the information rate with feedback design \mathbf{f} such that $\frac{1}{T+1} \mathbf{E} \|\underline{u}^T(\mathbf{f})\|^2 \leq \mathcal{P}$.

Alternatively

(T1') Optimal Feedback Communication Tradeoff: Given the channel \mathcal{F} with one-step delayed output feedback and an information rate \mathcal{R} , the achievable average channel input power $P_T(\mathbf{f}, \mathcal{R})$ cannot be lower than a constant $P_{T,T}(\mathcal{R})$ for any feedback communication design; here $P_T(\mathbf{f}, \mathcal{R}) := \frac{1}{T+1} \mathbf{E} \|\underline{u}^T(\mathbf{f})\|^2$ is the average channel input power with feedback design \mathbf{f} such that $\frac{1}{T+1} I(\underline{u}^T(\mathbf{f}) \rightarrow \underline{y}^T(\mathbf{f})) \geq \mathcal{R}$.

Note that the average input power depends on the strictly causal feedback from the channel output; the information rate, however, is independent of the causal feedback, may be achieved by anti-causally processing the channel outputs \underline{y}^T , and hence can be used as a measure of anti-causality of the system.

A fundamental tradeoff for the estimation problem over the channel \mathcal{F} is the causal estimation performance versus anti-causal estimation performance. Assume a process \underline{x}^T is passed through the channel \mathcal{F} and generates measurements \underline{y}^T . Let $\underline{W} := \mathbf{R}^{-1} \underline{x}^T$, where $\mathbf{R} := (\mathbf{K}_{\underline{x}}^{(T)})^{1/2}$ if $\mathbf{K}_{\underline{x}}^{(T)}$ is of full rank; otherwise \mathbf{R} is such that $\mathbf{K}_{\underline{W}}$ is of full rank with $\text{rank}(\mathbf{K}_{\underline{W}}) = \text{rank}(\mathbf{K}_{\underline{x}}^{(T)})$ and $\mathbf{K}_{\underline{W}} = \mathbf{I}$. That is, \underline{W} may be viewed as the to-be-estimated, normalized signal that completely determines the process \underline{x}^T . Therefore we have a linear model $\underline{y}^T = \mathbf{Z}_T^{-1} \mathbf{R} \underline{W} + \underline{N}^T$. Again one can define innovation as $e_t := \bar{y}_t - \mathbf{E}(\bar{y}_t | \underline{y}^{t-1})$ for each t .

(T2) Optimal Estimation Tradeoff: Given the channel \mathcal{F} and the time-averaged one-step prediction MMSE

$$\text{PMMSE}_r := \frac{1}{T+1} \sum_{t=0}^T (r_t - \mathbf{E}(r_t | \underline{y}^{t-1}))^2, \quad (85)$$

the decay rate of the anti-causal, smoothing MMSE

$$\frac{1}{2(T+1)} \log \det \text{MMSE}_{\underline{W}}^{-1} = -\frac{1}{2(T+1)} \log \det \mathbf{E} (\underline{W} - \mathbf{E}(\underline{W} | \underline{y}^T)) (\underline{W} - \mathbf{E}(\underline{W} | \underline{y}^T))' \quad (86)$$

cannot be larger than a constant, and the average of innovations variance in log scale $\frac{1}{2(T+1)} \sum_{t=0}^T \log K_{e,t}$ cannot be larger than a constant, for any one-step predictor design and smoother design.

Alternatively

(T2') Optimal Estimation Tradeoff: Given the channel \mathcal{F} and the decay rate of the anti-causal, smoothing MMSE $\frac{1}{2(T+1)} \log \det \text{MMSE}_{\underline{W}}^{-1}$ (or the average of innovations variance in log scale $\frac{1}{2(T+1)} \sum_{t=0}^T \log K_{e,t}$), the time-averaged one-step prediction MMSE PMMSE_r cannot be smaller than a constant, for any one-step predictor design and smoother design.

Note that the prediction MMSE depends on causality, while the smoothing MMSE is anti-causal and independent of the causal processing (if any) done by the estimator. That is, this tradeoff is concerned with prediction versus smoothing tradeoff, or more fundamentally, the causality versus anti-causality tradeoff.

A fundamental tradeoff for the cheap control problem over the channel \mathcal{F} is the control performance (regulated output variance, in this case the variance of the channel input signal) versus the Bode integral (or the disturbance rejection measure, degree of anti-causality, as defined in (79)). View the channel input $u_t(\mathbf{f})$ as the regulated output with control design \mathbf{f} , $y_t(\mathbf{f})$ be the associated channel output, and

$$\text{BI}_T(\mathbf{f}) := \frac{1}{2} \sum_{i=1}^{T+1} \log \lambda_i(\mathbf{K}_{\underline{y}}^{(T)}(\mathbf{f})). \quad (87)$$

(T3) Optimal Feedback Control Tradeoff: Given the channel \mathcal{F} and the average regulated output variance $\frac{1}{T+1} \sum_{t=0}^T \mathbf{E}(u_t(\mathbf{f}))^2$, the Bode integral $\text{BI}_T(\mathbf{f})$ cannot be larger than a constant for any control design \mathbf{f} .

Alternatively

(T3') Optimal Feedback Control Tradeoff: Given the channel \mathcal{F} and the Bode integral, the average regulated output variance cannot be smaller than a constant for any control design \mathbf{f} .

Note that this specifies the relation between the control performance achievable via causal feedback and the anti-causality of the system (that is, the Bode sensitivity integral or disturbance rejection measure which is independent of causal feedback).

To summarize, we have seen that *all three tradeoffs are essentially the fundamental tradeoff between causality and anti-causality*, which manifests itself in the three different but closely related problems. The causal entities, e.g. the channel inputs in feedback communication, one-step prediction in estimation, and regulated output in control, are closed-loop entities generated in a causal, progressive way by the causal feedback, and hence vary as the causal feedback varies. On the other hand, the anti-causal entities, e.g. the information rate (and the decoded message) in communication, the smoothed estimate in estimation, and the BI in control, are invariant regardless of whether the systems are in open-loop or closed-loop or how the closed-loop is done. It is worth noting the various discussions involving causal versus anti-causal operations and filtering versus smoothing in the literature; see [24], [25] and therein references.

In contrast, the power versus rate tradeoff in communication problems without output feedback cannot be interpreted as causality versus anti-causality tradeoff, nor can the tradeoff in the corresponding estimation problems. To see this, we again assume the linear Gaussian model $\underline{y}^T = \mathbf{Z}_T^{-1} \mathbf{R} \underline{W} + \underline{N}^T$. One can see that the channel input power is related to the unknown's prior covariance (i.e. covariance matrix of channel input $\mathbf{R} \underline{W}$), whereas the mutual information $I(\mathbf{W}; \underline{y}^T) = \frac{1}{2} \log \det \text{MMSE}_{\underline{W}}^{-1}$ is related to the posterior covariance (cf. Theorem 10.3, [43]). Thus, in communication without output feedback, the power versus rate tradeoff may be translated into the tradeoff between the unknown's prior covariance and posterior covariance (or more generally the tradeoff between the unknown's prior and posterior distributions). Note it is easily verified that these two tradeoffs coincide in the AWGN channel case as one might expect.

VIII. NECESSARY CONDITIONS FOR THE OPTIMALITY OF THE FINITE-HORIZON CODING STRUCTURE \mathbb{S}

We discuss in this section a few useful properties of the coding structure \mathbb{S}^* with the optimal feedback generator. The first two properties, i.e., the orthogonality between future channel inputs and previous channel outputs, and the Gauss-Markov property of the transformed channel outputs, are direct consequences of the KF. Naturally, they can be viewed as necessary conditions for optimality of the feedback communication scheme as we have proven the necessity of the KF for optimality. The third property, the finite-dimensionality of the optimizing \underline{z}^T , yet another necessary condition for optimality, is a joint consequence of the KF structure and the waterfilling requirement for optimality for the finite-dimensional channel \mathcal{F} . Finally, we show that the MMSE one-step predictor is necessary for achieving the feedback capacity of general additive channels with an average power constraint, followed by an extension of the orthogonality property over such channels.

A. Necessary condition for optimality: Orthogonality condition

First, we show that the coding structure \mathbb{S}^* satisfies a necessary condition for optimality discussed in [15]¹⁵. The condition says that, the channel input u_t needs to be orthogonal to the past channel outputs \underline{y}^{t-1} . This is intuitive since to ensure the fastest transmission, the transmitter should not (re-)transmit any information that the receiver has already obtained, thus the transmitter needs to remove any correlation with \underline{y}^{t-1} in u_t (to this aim, the transmitter has to access the channel outputs through feedback). This property, albeit a rather natural/simple consequence due to the Kalman filter, can yield interesting results, see e.g. [33].

¹⁵This was later referred to as the orthogonality condition in [33], based on which a Kalman filter structure is identified. It was also discussed in [12], [45].

Proposition 4. *In system (56), for any $0 \leq \tau < t$, it holds that $\mathbf{E}u_t e_\tau = 0$ and $\mathbf{E}u_t y_\tau = 0$. Equivalently, matrices $\mathbf{E}\underline{u}^T \underline{y}^{T'}$ and $\mathbf{E}\underline{u}^T \underline{e}^{T'}$ are upper triangular for any T .*

The justification of this proposition follows simply from the famous Projection Theorem (for MMSE estimators, the estimation error at a time is orthogonal to all available measurements, see e.g. [23], [43], [46]) which holds for the KF. Here note that u_t is in fact the one-step prediction error (i.e. $u_t = \underline{C}' \tilde{x}_t$ where \tilde{x}_t is the estimation error for \underline{x}_t with one-step prediction using the estimator $\mathbf{E}(\underline{x}_t | \underline{y}^{t-1})$). We also provide an alternative proof based on the state-space model in the appendix.

Proof: See Appendix IV. □

B. Gauss-Markov property of the transformed output process

In this subsection, we show that the process $\underline{\check{y}}^T$, a transformation of the output process \underline{y}^T or $\underline{\check{y}}^T$, is a Gauss-Markov (GM) process. In particular, it is an MA- m Gaussian process. This is a generalization of the result obtained in [9], which states that if the channel has an MA- m Gaussian noise process and has no ISI, a necessary condition for optimality is that the channel output needs to be an MA- m Gaussian process; see Corollary IV.1 in [33] for the detailed statement and proof of the result of [9]. This result has been generalized in [33], that is, if the channel has an m th order autoregressive moving-average (ARMA- m) Gaussian process and has no ISI, a necessary condition for optimality is that the channel output needs to be an ARMA- m Gaussian process; see Proposition VII.1 in [33]. Our result here, on the other hand, is concerned with an transformed output which is sometimes simpler to deal with.

Recall the relevant definitions in (21) and (22) of Section III, and define the transformed output process

$$\underline{\check{y}}^T := \mathbf{Z}_{z,T} \underline{y}^T. \quad (88)$$

From (22), it holds that

$$\underline{\check{y}}^T = \mathbf{Z}_{z,T} (\mathbf{Z}_T^{-1} \underline{u}^T + \underline{N}^T) = \mathbf{Z}_{p,T} \underline{u}^T + \mathbf{Z}_{z,T} \underline{N}^T. \quad (89)$$

This implies that, \check{y}_{t+m+1} is a linear combination of $\underline{u}_{t+1}^{t+m+1}$, and \check{y}_t is a linear combination of \underline{u}_{t-m}^t , since $\mathbf{Z}_{p,T}$ is banded (and lower triangular) with bandwidth $(m+1)$. But the Projection Theorem yields that $\underline{u}_{t+1}^{t+m+1}$ is independent of \check{y}_t , so \check{y}_{t+m+1} is independent of \check{y}_t . Repeat this argument and we can show that $\underline{\check{y}}^T$ is a banded process, i.e., an MA- m process. More formally, we have

Proposition 5. *In system (56), it holds that the transformed output process $\underline{\check{y}}^T$ is an MA- m Gaussian process, or equivalently*

$$\mathbf{K}_{\underline{\check{y}}}^{(T)} := \mathbf{E} \underline{\check{y}}^T \underline{\check{y}}^{T'} \quad (90)$$

is banded with bandwidth $(2m+1)$, i.e., $\mathbf{K}_{\underline{\check{y}}}^{(T)}(i, j) = 0$ if $|i - j| \geq m + 1$.

Proof: See Appendix IV. □

As a result of this proposition, we see that $\underline{\check{y}}^T$ is an ARMA- m process, as claimed in [33].

The different forms of channel outputs, i.e. \underline{y}^T , $\underline{\check{y}}^T$, and $\underline{\check{y}}^T$, causally determine each other; see Fig. 9 for their relations. Fig. 9 (a) shows the ISI-free colored Gaussian noise channel with a direct channel output $\underline{\check{y}}^T$ and a transformed output \underline{y}^T . Since this channel has no ISI, the optimal effective input process must waterfill the effective noise spectrum and hence $\underline{\check{y}}^T$ is the waterfilling output for the optimal scheme. Fig. 9 (b) shows the ISI channel corrupted by AWGN, with a channel output \underline{y}^T . Since the channel noise is white, it may be easy to directly apply the KF algorithm. Fig. 9 (c) is an ISI channel corrupted by a colored Gaussian noise with a channel output $\underline{\check{y}}^T$, but both the ISI filter and the filter generating the colored noise are MA- m filters. It may be easily used to establish that $\underline{\check{y}}^T$ is an MA- m process. These formulations are T -equivalent and can be easily converted from one to another.

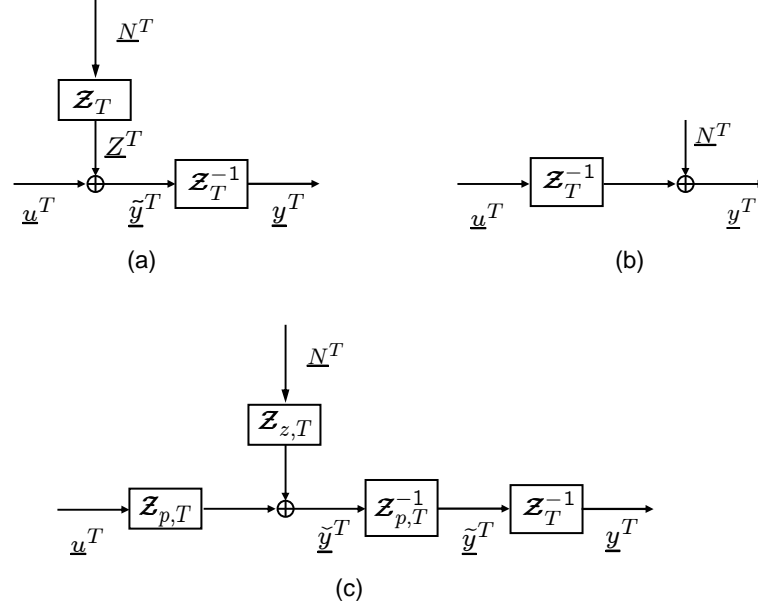


Fig. 9. (a) A colored Gaussian noise channel without ISI. This formulation may be directly used to study the waterfilling property of the optimal solution. (b) An equivalent ISI channel with AWGN. This formulation may be easily used to study the KF properties of the optimal solution. (c) Another equivalent channel model with both ISI and colored noise, but the ISI and colored noise filter are both MA- m filters. This formulation may be used to study the finite-dimensionality of the channel input/output processes. Note that \mathbf{Z}_T can be realized as $(\mathbf{F} - \underline{\mathbf{G}}\mathbf{H}', -\underline{\mathbf{G}}, \underline{\mathbf{H}}', 1)$, \mathbf{Z}_T^{-1} as $(\mathbf{F}, \underline{\mathbf{G}}, \underline{\mathbf{H}}', 1)$, $\mathbf{Z}_{p,T}$ as $(\mathbf{F}_z, \underline{\mathbf{G}}_p, \underline{\mathbf{H}}', 1)$, $\mathbf{Z}_{p,T}^{-1}$ as $(\mathbf{F} + \underline{\mathbf{G}}_z\mathbf{H}', \underline{\mathbf{G}}_p, \underline{\mathbf{H}}', 1)$, and $\mathbf{Z}_{z,T}$ as $(\mathbf{F}_z, \underline{\mathbf{G}}_z, \underline{\mathbf{H}}', 1)$

C. Finite dimensionality of the optimizing \underline{r}^T

We now show that, to achieve the finite-horizon feedback capacity $C_{T,n}$, the covariance matrix of the feedback-free, message-carrying process \underline{r}^T can have rank at most $(m+1)$, where m is the order of the channel $\mathcal{Z}(z)$. This is an extension of the finite-rankness property by Ordentlich (c.f. [9], [33]) for a Gaussian channel with an MA- m noise process to a Gaussian channel with an ARMA- m noise process.

Proposition 6. *For system (56), the optimal $\mathbf{K}_{\underline{r}}^{(T)}$ that solves $C_{T,n}$ as defined in (1) has rank at most $(m+1)$.*

The proof of this proposition is based on Lemma 2 below. This lemma deals with a special class of m th order channel \mathbf{Z}_T , that is, any \mathbf{Z}_T such that $(f_0 + g_0) = 0$ (see Sec. III for notations). In other words, $\mathbf{Z}_{p,T}$ is in fact an MA- $(m-1)$ model. For this class of channels, it is easy to extend the idea of Ordentlich (c.f. [9]) and prove that the optimal $\mathbf{K}_{\underline{r}}^{(T)}$ has rank at most m . Then the proposition can be proven by approaching any arbitrary \mathbf{Z}_T by elements in the special class of channels based on certain continuity properties.

Lemma 2. *For system (56) with $(f_0 + g_0) = 0$, the optimal $\mathbf{K}_{\underline{r}}^{(T)}$ that solves $C_{T,n}$ as defined in (1) has rank at most m .*

See Appendix IV-A for the proofs of the lemma and the proposition.

D. Necessity of the MMSE predictor for general channels with feedback

The necessity of the Kalman filter in achieving the optimality for the channel \mathcal{F} under an average power constraint can be easily extended. Assume an arbitrary additive channel

$$\underline{y}^T = \mathbf{H}\underline{u}^T + \underline{z}^T \quad (91)$$

with an average power constraint $\mathbf{E}\|\underline{u}^T\|^2 \leq (T+1)\mathcal{P}$, where \underline{Z}^T is an arbitrary additive noise process. Assuming one-step delayed channel output feedback, then such a channel needs to contain an MMSE one-step predictor in order to achieve the feedback capacity.

Proposition 7. *Let $\tilde{u}_t := u_t - \mathbf{E}(u_t|\underline{y}^{t-1})$ and $\tilde{\underline{y}}^T := \mathbf{H}\tilde{\underline{u}}^T + \underline{Z}^T$. Then $I(\underline{u}^T \rightarrow \underline{y}^T) = I(\tilde{\underline{u}}^T \rightarrow \tilde{\underline{y}}^T)$ and $\mathbf{E}\underline{u}^{T'}\underline{u}^T \geq \mathbf{E}\tilde{\underline{u}}^{T'}\tilde{\underline{u}}^T$.*

Proof: Note that $\mathbf{E}(u_t|\underline{y}^{t-1})$ can be generated and added back to the channel output at the receiver side and hence the directed information across the channel or mutual information from the message to channel outputs is the same using either \underline{u}^T or $\tilde{\underline{u}}^T$ as the channel inputs. The average power of using $\tilde{\underline{u}}^T$ is no larger since it has minimum variance. \square

Simple as it is, this necessary condition for optimality is rather universal. A corollary is that in the optimal feedback coding scheme the current channel input u_t is independent of all past channel outputs \underline{y}^{t-1} by the Projection Theorem, an extension of Proposition 4. Moreover, since $\mathbf{E}\tilde{\underline{u}}^T = 0$ by the law of total expectation, it is a center-of-gravity encoding rule (cf. [12], [37]). It is also straightforward to see that if the channel output feedback delay is d steps, then an MMSE d -step predictor is needed for optimality.

IX. ASYMPTOTIC ANALYSIS OF THE FEEDBACK SYSTEM

By far we have completed our analysis in finite-horizon. We have shown that the optimal design of encoder and decoder must contain a KF, and connected the feedback communication problem to an estimation problem and a control problem. Below, we briefly consider the steady-state communication problem, by studying the limiting behavior (T going to infinity) of the finite-horizon solution while fixing the encoder dimension to be $(n+1)$. The infinite-horizon capacity problem will not be considered in this paper. Here and hereafter, we make the following assumption unless otherwise specified:

(A2): $(\mathbf{A}, \underline{\mathbf{C}}')$ is observable, and none of the eigenvalues of \mathbf{A} are on the unit circle or at the locations of the eigenvalues of \mathbf{F} .

A. Convergence to steady-state

The time-varying KF in (62) converges to a steady-state, namely (62) is *stabilized* in closed-loop: The distributions of u_t , e_t , and y_t will converge to steady-state distributions, and $\underline{\Sigma}_t$, \underline{L}_t , $\underline{\mathbf{G}}_t^*(\mathbf{A}, \underline{\mathbf{C}})$, $\hat{\underline{\mathbf{G}}}_t^*$, and $K_{e,t}$ will converge to their steady-state values. That is, asymptotically (62) becomes an LTI system

$$\text{steady-state: } \begin{cases} \underline{\mathbf{X}}_{t+1} &= (\mathbf{A} - \underline{\mathbf{L}}\underline{\mathbf{C}}')\underline{\mathbf{X}}_t - \underline{\mathbf{L}}N_t = \mathbf{A}\underline{\mathbf{X}}_t - \underline{\mathbf{L}}e_t \\ e_t &= \underline{\mathbf{C}}'\underline{\mathbf{X}}_t + N_t \\ u_t &= \underline{\mathbf{D}}'\underline{\mathbf{X}}_t, \end{cases} \quad (92)$$

where

$$\underline{\mathbf{L}} := \frac{\mathbf{A}\underline{\Sigma}\underline{\mathbf{C}}}{K_e}, \quad (93)$$

$K_e = \underline{\mathbf{C}}'\underline{\Sigma}\underline{\mathbf{C}} + 1$, and $\underline{\Sigma}$ is the unique stabilizing solution to the Riccati equation

$$\underline{\Sigma} = \mathbf{A}\underline{\Sigma}\mathbf{A}' - \frac{\mathbf{A}\underline{\Sigma}\underline{\mathbf{C}}\underline{\mathbf{C}}'\underline{\Sigma}\mathbf{A}'}{\underline{\mathbf{C}}'\underline{\Sigma}\underline{\mathbf{C}} + 1}. \quad (94)$$

This LTI system is sometimes easy to analyze (e.g., it allows transfer function based study) and to implement. For instance, the cheap control (cf. [21] and [44]) of an LTI system claims that the transfer function from N to e is an *all-pass* function in the form of

$$\mathcal{T}_{Ne}(z) = \prod_{i=0}^k \frac{z - \lambda_i}{z - \lambda_i^{-1}} \quad (95)$$

where $\lambda_0, \dots, \lambda_k$ are the *unstable eigenvalues* of \mathbf{A} or \mathbf{A} (noting that \mathbf{F} is stable). Note that this is consistent with the whiteness of innovations process $\{e_t\}$.

The existence of steady-state of the KF is proven in the following proposition. Notice that (62) is a *singular* KF since it has no process noise; the convergence of such a problem was established in [47].

Proposition 8. *Consider the Riccati recursion (65) and the system (62). Assume (A2) and that $\lambda_0, \dots, \lambda_k$ are the unstable eigenvalues of \mathbf{A} .*

- i) *Starting from the initial condition given in (66), the Riccati recursion (65) generates a sequence $\{\Sigma_t\}$ that converges to Σ_∞ , the unique stabilizing solution to the Riccati equation (94), and Σ_∞ has rank $(k+1)$.*
- ii) *The time-varying system (62) converges to the unique steady-state as given in (92).*

Proof: See Appendix V. □

B. Steady-state quantities

Now we fix $(\mathbf{A}, \underline{\mathbf{C}})$ and let the horizon T in the coding structure \mathbb{S}^* go to infinity. Let $\mathcal{H}(e)$ be the entropy rate of $\{e_t\}$,

$$DI(\mathbf{A}) := \prod_{i=0}^k |\lambda_i| \quad (96)$$

be the *degree of instability* or the *degree of anti-causality* of \mathbf{A} , and $S(e^{j2\pi\theta}) := Y(e^{j2\pi\theta})/N(e^{j2\pi\theta})$ be the spectrum of the sensitivity function of system (92) (cf. [14]). Then the limiting result of Proposition 3 is summarized in the next proposition.

Proposition 9. *Consider the coding structure \mathbb{S}^* . For any $n \geq 0$ and $(\mathbf{A}, \underline{\mathbf{C}}')$ with $\mathbf{A} \in \mathbb{R}^{(n+1) \times (n+1)}$ satisfying (A2),*

- i) *The asymptotic information rate is given by*

$$\begin{aligned} R_{\infty,n}(\mathbf{A}, \underline{\mathbf{C}}) &:= \lim_{T \rightarrow \infty} \frac{1}{T+1} I(\underline{\mathbf{W}}; \underline{\mathbf{y}}^T) \\ &= \mathcal{H}(e) - \frac{1}{2} \log 2\pi e \\ &= \frac{1}{2} \log K_e \\ &= \log DI(\mathbf{A}) \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \log |S(e^{j2\pi\theta})| d\theta \\ &= \frac{1}{2} \log(\underline{\mathbf{C}}' \Sigma \underline{\mathbf{C}} + 1) \\ &= \lim_{T \rightarrow \infty} \frac{\log \det \mathbf{I}_{\underline{\mathbf{W}}, T}}{2(T+1)} \\ &= - \lim_{T \rightarrow \infty} \frac{\log \det \mathbf{MSE}_{\underline{\mathbf{W}}, T}}{2(T+1)} \\ &= - \lim_{T \rightarrow \infty} \frac{\log \det \mathbf{CRB}_{\underline{\mathbf{W}}, T}}{2(T+1)}. \end{aligned} \quad (97)$$

- ii) *The average channel input power is given by*

$$\begin{aligned} P_{\infty,n}(\mathbf{A}, \underline{\mathbf{C}}) &:= \lim_{T \rightarrow \infty} \frac{1}{T+1} \mathbf{E} \|\underline{\mathbf{u}}^T\|^2 \\ &= \underline{\mathbf{D}}' \Sigma \underline{\mathbf{D}}. \end{aligned} \quad (98)$$

Remark 6. Proposition 9 links the asymptotic information rate to the degree of anti-causality and Bode sensitivity integral ([14]) for the control system, to the entropy rate and steady-state variance of the innovations process, asymptotic increasing rate of the Fisher information, and the asymptotic decay rate of smoothing MSE or of CRB for the estimation system. Note that the Bode sensitivity integral is the fundamental limitation of the disturbance rejection (control) problem, and the asymptotic decay rate of CRB is the fundamental limitation of the recursive estimation problem. Hence, the fundamental limitations in feedback communication, control, and estimation coincide. More specifically, the asymptotic information rate cannot be made higher or lower than a constant regardless of the feedback generator choice; the disturbance rejection measure cannot be made smaller than a constant regardless of the feedback controller design; the decay rate of the estimate error cannot be made faster than a constant regardless of the estimator design; and the constant is the logarithm of the degree of anti-causality of \mathbf{A} .

Remark 7. It is straightforward to extend the finite-horizon connections between the fundamental tradeoffs for feedback communication, estimation, and feedback control to infinite horizon. As the limits exist, quantities in fundamental tradeoffs (T1) through (T3) given in Section VII-C are well defined in infinite horizon and the corresponding relationship still holds. Note it is more obvious to see that the Bode integral is associated with anti-causality since it equals the logarithm of the degree of anti-causality of \mathbf{A} .

Proof: Proposition 8 leads to that, the limits of the results in Proposition 3 are well defined. Then

$$\begin{aligned} R_{\infty,n}(\mathbf{A}, \underline{\mathcal{C}}) &= \lim_{T \rightarrow \infty} \frac{1}{2(T+1)} \sum_{t=0}^T \log K_{e,t} \\ &= \lim_{T \rightarrow \infty} \frac{1}{2} \log K_{e,T} \\ &= \mathcal{H}(e) - \frac{1}{2} \log 2\pi e, \end{aligned} \tag{99}$$

where the second equality is due to the Cesaro mean (i.e., if a_k converges to a , then the average of the first k terms converges to a as k goes to infinity), and the last equality follows from the definition of entropy rate of a Gaussian process (cf. [36]).

Now by (95), $\{e_t\}$ has a flat power spectrum with magnitude $DI(\mathbf{A})^2$. Then $R_{\infty,n}(\mathbf{A}, \underline{\mathcal{C}}) = \log DI(\mathbf{A})$. The Bode integral of sensitivity follows from [14]. The other equalities are the direct applications of the Cesaro mean to the results in Proposition 3. \square

Proposition 9 implies that the presence of stable eigenvalues in \mathbf{A} does not affect the rate (see also [14]). Stable eigenvalues do not affect $P_{\infty,n}(\mathbf{A}, \underline{\mathcal{C}})$, either, since the initial condition response associated with the stable eigenvalues can be tracked with zero power (i.e. zero average MSE). Therefore, we conclude that the presence of stable eigenvalues in \mathbf{A} does not affect either the rate $R_{\infty,n}(\mathbf{A}, \underline{\mathcal{C}})$ or the power $P_{\infty,n}(\mathbf{A}, \underline{\mathcal{C}})$. We have thus seen that the communication problem is essentially a problem of tracking an anti-causal source over a communication channel ([13], [14], [20]).

Corollary 1. Suppose that $(\mathbf{A}, \underline{\mathcal{C}})$ with $\mathbf{A} \in \mathbb{R}^{(n+1) \times (n+1)}$ satisfies (A2). Suppose further that \mathbf{A} has $(k+1)$ unstable eigenvalues denoted $\lambda_0, \dots, \lambda_k$ where $0 \leq k \leq (n+1)$. Then there exists an observable pair $(\mathbf{A}_k, \underline{\mathcal{C}}'_k)$ with $\mathbf{A}_k \in \mathbb{R}^{(k+1) \times (k+1)}$ being anti-stable such that $R_{\infty,n}(\mathbf{A}, \underline{\mathcal{C}}) = R_{\infty,k}(\mathbf{A}_k, \underline{\mathcal{C}}_k)$ and $P_{\infty,n}(\mathbf{A}, \underline{\mathcal{C}}) = P_{\infty,k}(\mathbf{A}_k, \underline{\mathcal{C}}_k)$.

Proof: See Appendix VI. \square

X. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a perspective that integrates information transmission (communication), information processing (estimation), and information utilization (control). We identified and explored fundamental limitations in feedback communication, estimation, and feedback control over Gaussian channels with memory. Specifically, we

established a certain equivalence of a feedback communication system, an estimation system, and a feedback control system. We demonstrated that a simple reformulation of the Kalman filter becomes the celebrated Schalkwijk-Kailath codes, and the well-studied Cover-Pombra structure necessarily contains a Kalman filter in order to be optimal. We characterized the roles of Kalman filtering in an optimal feedback communication system as to ensure power efficiency and to optimally recover the transmitted codewords. We showed that the fundamental limitations/tradeoffs in these three systems also coincide: The power versus rate tradeoff in feedback communication, the causal prediction versus smoothing tradeoff in estimation, and the control performance versus Bode integral tradeoff in control, are equivalent and in essence, all of them are the causality versus anti-causality tradeoffs. We also presented a coding scheme achieving the finite-horizon feedback capacity of the Gaussian channel. The scheme is based on the Kalman filtering algorithm, and provides refinements and extensions to the Cover-Pombra coding structure and Schalkwijk-Kailath codes.

Our new perspective has been recently generalized in [22] to uniformly address the fundamental limits of several classes of feedback communication problems, and we envision that this perspective can generate a new avenue for studying more general feedback communication problems, such as multiuser feedback communications. Our ongoing research includes extending our proposed scheme to address the optimality of more feedback communication problems (such as single-user MIMO systems with output feedback, multi-user MIMO systems with output feedback). We also anticipate that the perspective and the approaches developed in this paper be extended and help to build a theoretically and practically sound paradigm that unifies information, estimation, and control.

APPENDIX I

SYSTEMS REPRESENTATIONS AND EQUIVALENCE

The concept of system representations and the equivalence between different representations are extensively used in this paper. In this subsection, we briefly introduce system representations and the equivalence. For more thorough treatment, see e.g. [48]–[50].

A. Systems representations

Any discrete-time linear system can be represented as a linear mapping (or a linear operator) from its input space to output space; for example, we can describe a single-input single-output (SISO) linear system as

$$\underline{y}^t = \mathbf{M}_t \underline{u}^t \quad (100)$$

for any t , where $\mathbf{M}_t \in \mathbb{R}^{(t+1) \times (t+1)}$ is the matrix representation of the linear operator, $\underline{u}^t \in \mathbb{R}^{t+1}$ is the stacked input vector consisting of inputs from time 0 to time t , and $\underline{y}^t \in \mathbb{R}^{t+1}$ is the stacked output vector consisting of outputs from time 0 to time t . For a (strictly) causal SISO LTI system, \mathbf{M}_t is a (strictly) lower triangular Toeplitz matrix formed by the coefficients of the impulse response. Such a system may also be described as the (reduced) transfer function, whose inverse z -transform is the impulse response; by a (reduced) transfer function we mean that its zeros are not at the same location of any pole.

A causal SISO LTI system can be realized in state-space as

$$\begin{cases} \underline{x}_{t+1} &= \mathbf{A} \underline{x}_t + \mathbf{B} u_t \\ y_t &= \underline{\mathbf{C}}' \underline{x}_t + D u_t, \end{cases} \quad (101)$$

where $\underline{x}_t \in \mathbb{R}^l$ is the state, $u_t \in \mathbb{R}$ is the input, $y_t \in \mathbb{R}$ is the output, \underline{A} is the state matrix, \underline{B} is the input matrix (vector), \underline{C} is the output matrix (vector), and D is the direct feedthrough term. We call l the *dimension* or the *order* of the realization. The state-space representation (101) may be conveniently denoted as $(\underline{A}, \underline{B}, \underline{C}', D)$. Note that in the study of input-output relations, it is sometimes convenient to assume that the system is relaxed or at initial rest (i.e. zero input leads to zero output), whereas in the study of state-space, we generally allow $\underline{x}_0 \neq \underline{0}$, which is not at initial rest. For multi-input multi-output (MIMO) systems, linear time-varying systems, etc., see [49], [50].

The state-space representation of an causal FDLTI system $\mathcal{M}(z)$ is not unique. We call a realization $(\underline{A}, \underline{B}, \underline{C}', D)$ *minimal* if $(\underline{A}, \underline{B})$ is controllable and $(\underline{A}, \underline{C}')$ is observable. All minimal realizations of $\mathcal{M}(z)$ have the same dimension, which is the minimum dimension of all possible realizations. All other realizations are called *non-minimal*. The transfer function for the state-space representation $(\underline{A}, \underline{B}, \underline{C}', D)$ is $\underline{C}'(z\mathbf{I} - \underline{A})^{-1}\underline{B} + D$.

Example: Derivation of state-space representation of $\mathcal{G}_T^*(\underline{A}, \underline{C})$

We demonstrate here how we can derive a realization of a system. Consider $\mathcal{G}_T^*(\underline{A}, \underline{C})$ in (49) in Section V, which is given by

$$\mathcal{G}_T^*(\underline{A}, \underline{C}) = -\hat{\mathcal{G}}_T^*(\mathbf{I} - \mathbf{Z}_T^{-1}\hat{\mathcal{G}}_T^*)^{-1}, \quad (102)$$

where the state-space representations for $\hat{\mathcal{G}}_T^*(\underline{A}, \underline{C})$ and \mathbf{Z}_T^{-1} are illustrated in Fig. 8 (b) and Fig. 3 (c). This result shows that the block diagram in Fig. 6 (c) is indeed the dynamics of $\hat{\mathcal{G}}_T^*$, as claimed in Proposition 2 iii).

Since (102) suggests a feedback connection of $\hat{\mathcal{G}}^*$ and \mathbf{Z}^{-1} as shown in Fig. 10, we can write the state-space for \mathcal{G}^* as

$$\left\{ \begin{array}{l} \hat{\underline{x}}_{t+1} = \underline{A}\hat{\underline{x}}_t + \underline{L}_{1,t}e_t \\ \hat{r}_t = \underline{C}'\hat{\underline{x}}_t \\ \hat{\underline{s}}_{t+1} = \underline{F}\hat{\underline{s}}_t + \underline{G}\hat{r}_t + \underline{L}_{2,t}e_t \\ e_t = \underline{y}_t - \underline{H}'\hat{\underline{s}}_t - \hat{r}_t \\ \underline{s}_{a,t+1} = \underline{F}\underline{s}_{a,t} + \underline{G}\hat{r}_t \\ \underline{y}_t = y_t + \underline{H}'\underline{s}_{a,t} + \hat{r}_t \end{array} \right\} \begin{array}{l} \text{Kalman filter } \hat{\mathcal{G}}_T^*(\underline{A}, \underline{C}) \\ \mathbf{Z}_T^{-1} \end{array} \quad (103)$$

Letting $\hat{\underline{s}}_t := \hat{\underline{s}}_t - \underline{s}_{a,t}$, the above reduces to

$$\left\{ \begin{array}{l} \hat{\underline{x}}_{t+1} = \underline{A}\hat{\underline{x}}_t + \underline{L}_{1,t}e_t \\ \hat{r}_t = \underline{C}'\hat{\underline{x}}_t \\ \hat{\underline{s}}_{t+1} = \underline{F}\hat{\underline{s}}_t + \underline{L}_{2,t}e_t \\ e_t = y_t - \underline{H}'\hat{\underline{s}}_t. \end{array} \right. \quad (104)$$

This is the dynamics shown in Fig. 6 (c). Note that the above reduction of realization is allowed since it preserves the “ T -equivalence”, see the next subsection.

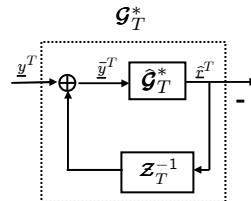


Fig. 10. \mathcal{G}^* is a feedback connection of $\hat{\mathcal{G}}^*$ and \mathbf{Z}^{-1} .

Example: State-space representation of an inverse of a system

Given a linear system with input u and output y and represented as $(\underline{A}, \underline{B}, \underline{C}', 1)$ in state space, it can be inverted if it is both stable and minimum-phase. The inverse system that maps y back to u can be realized in state space as $(\underline{A} - \underline{B}\underline{C}', -\underline{B}, \underline{C}', 1)$.

Example: State-space representations of \mathcal{Z}_z and \mathcal{Z}_p

It is easily shown that $\mathcal{Z}_z(z)$ can be realized as $(\underline{F}_z, \underline{G}_z, \underline{H}', 1)$, $\mathcal{Z}_p(z)$ can be realized as $(\underline{F}_z, \underline{G}_p, \underline{H}', 1)$, and $\mathcal{Z}_z^{-1}(z)$ can be realized as $(\underline{F}, -\underline{G}_z, \underline{H}', 1)$, where $\underline{F}_z := \underline{F} + \underline{G}_z \underline{H}'$.

B. Equivalence between representations

Definition 2. i) Two FDLTI systems represented in state-space are said to be equivalent if they admit a common transfer function (or a common transfer function matrix) and they are both stabilizable and detectable.

ii) Fix $0 \leq T < \infty$. Two linear mappings $\mathcal{M}_{i,T} : \mathbb{R}^{q(T+1)} \rightarrow \mathbb{R}^{p(T+1)}$, $i = 1, 2$, are said to be T -equivalent if for any $\underline{u}^T \in \mathbb{R}^{q(T+1)}$, it holds that

$$\mathcal{M}_{1,T}(\underline{u}^T) = \mathcal{M}_{2,T}(\underline{u}^T). \quad (105)$$

We note that i) is defined for FDLTI systems, whereas ii) is for general linear systems. i) implies that, the realizations of a transfer function are not necessarily equivalent. However, if we focus on all realizations that do not “hide” any unstable modes, namely all the unstable modes are either controllable from the input or observable from the output, they are equivalent; the converse is also true. ii) concerns about the *finite-horizon* input-output relations only. Since the states are not specified in ii), it is not readily extended to infinite horizon: Any unstable modes “hidden” from the input and output will grow unboundedly regardless of input and output, which is unwanted.

Example: T -equivalence between the estimation system (57) and coding structure \mathbb{S} (56)

To show the T -equivalence, it is sufficient to show that for each t , the sets of signals $r_t, \hat{r}_t, e_t, \underline{x}_t$, and $\hat{\underline{x}}_t$ in (57) and (56) are equal, respectively. To this aim, first note that for $t = 0$, the sets signals are equal, respectively, and that $\underline{s}_0 - \hat{\underline{s}}_0 = \bar{\underline{s}}_0 - \hat{\bar{\underline{s}}}_0$. Assume that for $t \leq \tau$, the sets of signals are equal, respectively, and that $\underline{s}_\tau - \hat{\underline{s}}_\tau = \bar{\underline{s}}_\tau - \hat{\bar{\underline{s}}}_\tau$. Now use induction. Apparently, $r_{\tau+1}$ and $\underline{x}_{\tau+1}$ generated by (57) and (56) are equal, respectively. Then

$$\begin{aligned} \underline{s}_{\tau+1} - \hat{\underline{s}}_{\tau+1} &= \underline{F}(\underline{s}_\tau - \hat{\underline{s}}_\tau) + \underline{G}(r_\tau - \hat{r}_\tau) - \underline{L}_{2,\tau} e_\tau \\ &= \bar{\underline{s}}_{\tau+1} - \hat{\bar{\underline{s}}}_{\tau+1}, \end{aligned} \quad (106)$$

and $e_{\tau+1}$ from both (57) and (56) equals

$$\underline{H}'(\bar{\underline{s}}_\tau - \hat{\bar{\underline{s}}}_\tau) + (r_\tau - \hat{r}_\tau) + N_{\tau+1}. \quad (107)$$

Thus we have proven the T -equivalence.

Likewise, we can show that the estimation system (57), feedback communication system (56), and control system (62) are T -equivalent.

Examples

As we mentioned in Section III-B, for any \underline{u}^T and \underline{N}^T , Fig. 3 (a) and (b) generate the same channel output $\tilde{\underline{y}}^T$. That is, the mappings from $(\underline{u}^T, \underline{N}^T)$ to $\tilde{\underline{y}}^T$ for the two channels are identical, and both are given by

$$\tilde{\underline{y}}^T = \mathcal{Z}_T(\mathcal{Z}_T^{-1} \underline{u}^T + \underline{N}^T). \quad (108)$$

Thus, we say the two channels are T -equivalent.

APPENDIX II

INPUT/OUTPUT CHARACTERIZATION OF FINITE-HORIZON INFORMATION CAPACITY: DIRECTED INFORMATION

Definition 3. The directed information from \underline{u}^T to \underline{y}^T is defined as

$$I(\underline{u}^T \rightarrow \underline{y}^T) := \sum_{t=0}^T I(\underline{u}^t; y_t | \underline{y}^{t-1}). \quad (109)$$

See [11] for details. One important feature about the directed information is that it is an input-output counterpart of the mutual information, which is especially useful to deal with channels within loops.

Proposition 10. (Tatikonda and Mitter) It holds that

$$C_T(\mathcal{P}) = \sup_{\underline{u}^T} \frac{1}{T+1} I(\underline{u}^T \rightarrow \underline{y}^T), \quad (110)$$

where the supremum is over all possible feedback-dependent Gaussian input distributions satisfying the power constraint

$$\frac{1}{T+1} \mathbf{E} \underline{u}^{T'} \underline{u}^T \leq \mathcal{P}, \quad (111)$$

and in the form

$$u_t = \underline{\gamma}'_t \underline{u}^{t-1} + \underline{\eta}'_t y^{t-1} + \xi_t \quad (112)$$

for any $\underline{\gamma}_t \in \mathbb{R}^t$, $\underline{\eta}_t \in \mathbb{R}^t$, and zero-mean Gaussian random variable $\xi_t \in \mathbb{R}$ independent of \underline{u}^{t-1} and y^{t-1} .

This proposition follows directly from the following lemma.

Lemma 3. The CP structure for the ISI Gaussian channel \mathcal{F} shown in Fig. 4 can generate any Gaussian channel input process $\{u_t\}$ in the form of (112) and vice versa.

Proof: Note that any input generated by the scheme in Fig. 4 has the form of

$$\underline{u}^t = \mathbf{B}_t \mathbf{Z}_t \underline{N}^t + \underline{v}^t = \mathbf{B}_t \mathbf{Z}_t \underline{y}^t - \mathbf{B}_t \underline{u}^t + \underline{v}^t, \quad (113)$$

leading to

$$\underline{u}^t = (\mathbf{I} + \mathbf{B}_t)^{-1} \mathbf{B}_t \mathbf{Z}_t \underline{y}^t + (\mathbf{I} + \mathbf{B}_t)^{-1} \underline{v}^t, \quad (114)$$

where \underline{v}^t is independent of \underline{N}^t and hence \underline{Z}^t .

On the other hand, from (112), we have

$$\underline{u}^t = \bar{\boldsymbol{\gamma}}_t \underline{u}^t + \bar{\boldsymbol{\eta}}_t y^t + \underline{\xi}^t, \quad (115)$$

leading to

$$\underline{u}^t = (\mathbf{I} - \bar{\boldsymbol{\gamma}}_t)^{-1} \bar{\boldsymbol{\eta}}_t y^t + (\mathbf{I} - \bar{\boldsymbol{\gamma}}_t)^{-1} \underline{\xi}^t, \quad (116)$$

where $\bar{\boldsymbol{\gamma}}_t \in \mathbb{R}^{(t+1) \times (t+1)}$ is the strictly lower triangular matrix formed by $\underline{\gamma}'_0, \dots, \underline{\gamma}'_t$, and $\bar{\boldsymbol{\eta}}_t \in \mathbb{R}^{(t+1) \times (t+1)}$ is the strictly lower triangular matrix formed by $\underline{\eta}'_0, \dots, \underline{\eta}'_t$. Since for any $\tau \leq t$, ξ_τ is independent of $\underline{u}^{\tau-1}$ and $y^{\tau-1}$, ξ_τ is independent of $\underline{N}^{\tau-1}$. By causality ξ_τ is independent of $N_\tau, N_{\tau+1}, \dots$. Therefore, $\underline{\xi}^t$ is independent of \underline{N}^t and hence \underline{Z}^t . Then the lemma follows by comparing (114) and (116). \square

Proof of Proposition 10: This proposition follows trivially from the observation that, for any Gaussian input of form (33), it holds that

$$\begin{aligned}
I(\underline{u}^T \rightarrow \underline{y}^T) &= \sum_{t=0}^T (h(y_t | \underline{y}^{t-1}) - h(y_t | \underline{u}^t, \underline{y}^{t-1})) \\
&= h(\underline{y}^T) - \sum_{t=0}^T h(N_t | \underline{u}^t, \underline{y}^{t-1}) \\
&= h(\underline{y}^T) - h(\underline{N}^T) \\
&= I(\underline{r}^T; \underline{y}^T).
\end{aligned} \tag{117}$$

□

Note that the above proof can be easily used to show that $I(\underline{u}^T \rightarrow \underline{y}^T)$ equals $I(W; \underline{y}^T)$ and $I(\underline{\xi}^T; \underline{y}^T)$, where W is the message. That is, the directed information from the input signal to output signal (both signals are inside the feedback loop and causally affecting each other) effectively captures the mutual information between the message W (or a message-carrying signal outside the feedback loop, such as $\underline{\xi}^T$ or \underline{r}^T) and the output signal. Therefore, directed information has the advantage of capturing the capacity without the need to identify the message, see e.g. [12].

APPENDIX III

CORRESPONDING RELATION BETWEEN THE CP STRUCTURE AND CODING STRUCTURE §

i) Assume $\mathbf{K}_{\underline{r}}^{(T)} > 0$ first. For any fixed $(\mathbf{K}_{\underline{r}}^{(T)}, \mathbf{B}_T)$ in the CP structure, define in § that

$$\begin{aligned}
\mathcal{G}_T &:= (\mathbf{I} + \mathbf{B}_T)^{-1} \mathbf{B}_T \mathcal{Z}_T \\
\mathbf{A} &:= \Gamma_a^{-1} \left[\begin{array}{c|c} 0 & \mathbf{I}_T \\ \hline * & * \end{array} \right] \Gamma_a := \Gamma_a^{-1} \mathbf{A}_a \Gamma_a \in \mathbb{R}^{(T+1) \times (T+1)} \\
\mathbf{C} &:= \Gamma_a [1 \quad 0 \quad \cdots \quad 0]' := \Gamma_a \underline{e}_1,
\end{aligned} \tag{118}$$

where $\Gamma_a := (\mathbf{K}_{\underline{r}}^{(T)})^{\frac{1}{2}} = \Gamma'_a$ is a positive definite square root, $*$ can be any number, and \mathbf{A}_a and \underline{e}_1 are defined in obvious ways (note that this vector \underline{e}_1 should not be confused with the Kalman filter innovation e_t). Then it is easily verified that \mathcal{G}_T is strictly lower triangular and $(\mathbf{A}, \mathbf{C}')$ is observable¹⁶. One can compute that the observability matrix for $(\mathbf{A}, \mathbf{C}')$ is in fact Γ_a , that is,

$$\begin{bmatrix} \mathbf{C}' \\ \mathbf{C}' \mathbf{A} \\ \mathbf{C}' \mathbf{A}^2 \\ \vdots \\ \mathbf{C}' \mathbf{A}^T \end{bmatrix} = \begin{bmatrix} \underline{e}_1' \\ \underline{e}_1' \mathbf{A}_a \\ \underline{e}_1' \mathbf{A}_a^2 \\ \vdots \\ \underline{e}_1' \mathbf{A}_a^T \end{bmatrix} \Gamma_a = \Gamma_a, \tag{119}$$

where the last equality is due to the structures of \mathbf{A}_a and \underline{e}_1 . Thus, $(\mathbf{A}, \mathbf{C}')$ can generate process \underline{r}^T with covariance matrix $\mathbf{K}_{\underline{r}}^{(T)}$. Then by (33), we know that for any given $(\mathbf{K}_{\underline{r}}^{(T)}, \mathbf{B}_T)$ with $\mathbf{K}_{\underline{r}}^{(T)} > 0$, we can find an admissible $(\mathbf{A}, \mathbf{C}, \mathcal{G}_T)$ generating the same channel input \underline{u}^T as $(\mathbf{K}_{\underline{r}}^{(T)}, \mathbf{B}_T)$ does.

Now consider the case that $\mathbf{K}_{\underline{r}}^{(T)} \geq 0$ but $\mathbf{K}_{\underline{r}}^{(T)}$ is not positive definite. Consider the positive definite sequence $\{\mathbf{K}_{\underline{r}}^{(T)} + \frac{1}{i} \mathbf{I}\}_{i=1}^{\infty}$. Therefore, for each pair $(\mathbf{K}_{\underline{r}}^{(T)} + \frac{1}{i} \mathbf{I}, \mathbf{B}_T)$, we can find an admissible triple $(\mathbf{A}_i, \mathbf{C}_i, \mathcal{G}_{T,i})$ corresponding to it per above construction. It is easily shown that the sequence of triples generate a sequence of inputs that converge to $\underline{u}^T(\mathbf{K}_{\underline{r}}^{(T)}, \mathbf{B}_T)$. Note that, however, power constraint or rate constraint given in Definition 1 is not considered here and hence may not hold unless some further constraint on the sequence is imposed.

¹⁶ \mathbf{A} can also be chosen to be such that its eigenvalues are not on the unit circle and not at the locations of \mathbf{F} 's eigenvalues, as Assumption (A2) requires in order to guarantee convergence in Section IX.

ii) Conversely, for any fixed admissible $(\mathbf{A}, \underline{\mathbf{C}}, \mathbf{G}_T)$ with $\mathbf{A} \in \mathbb{R}^{(T+1) \times (T+1)}$, we can obtain an admissible $(\mathbf{K}_T^{(T)}, \mathbf{B}_T)$ as

$$\begin{aligned}\mathbf{B}_T &:= \mathbf{G}_T \mathbf{Z}_T^{-1} (\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1})^{-1} \\ \mathbf{K}_T^{(T)} &:= \mathbf{\Gamma}_T(\mathbf{A}, \underline{\mathbf{C}}) \mathbf{\Gamma}_T(\mathbf{A}, \underline{\mathbf{C}})',\end{aligned}\tag{120}$$

which generates identical channel input \underline{u}^T as $(\mathbf{A}, \underline{\mathbf{C}}, \mathbf{G}_T)$ does.

iii) By the continuity of the mutual information and power, the limits of the power sequence and the mutual information sequence generated by $\underline{u}^T(\mathbf{A}_i, \underline{\mathbf{C}}_i, \mathbf{G}_{T,i})$ are equal to the power and mutual information generated by $\underline{u}^T(\mathbf{K}_T^{(T)}, \mathbf{B}_T)$, respectively. Then note that $P_{T,T}(\mathcal{R})$ is the infimum power over all admissible $(\mathbf{A}, \underline{\mathbf{C}}, \mathbf{G}_T)$ with $\mathbf{A} \in \mathbb{R}^{(T+1) \times (T+1)}$ according to Definition 1, and that $P_T(\mathcal{R})$ is the infimum power over all admissible $(\mathbf{K}_T^{(T)}, \mathbf{B}_T)$ according to (29), subject to rate constraints in (36) and (42) respectively, which implies that to show iii), it is sufficient to show that the equivalent inputs (or equivalent input sequences) hold the rate constraint in Definition 1. By the construction of the sequence of admissible triples $\{(\mathbf{A}_i, \underline{\mathbf{C}}_i, \mathbf{G}_{T,i})\}_{i=1}^\infty$, it is straightforward to see that the rate constraint (i.e. $I(\underline{W}; \underline{y}^T)/(T+1) \geq \mathcal{R}$) can be satisfied, yielding $P_T(\mathcal{R}) = P_{T,T}(\mathcal{R})$ and hence $C_T(\mathcal{P}) = C_{T,T}(\mathcal{P})$.

On the other hand, one can directly prove $C_T(\mathcal{P}) = C_{T,T}(\mathcal{P})$ without resorting to $P_{T,T}(\mathcal{R})$. To this aim, use an arbitrarily small reduction $\epsilon > 0$ from the power budget \mathcal{P} , that is, consider only those $(\mathbf{K}_T^{(T)}, \mathbf{B}_T)_\epsilon$ such that $\mathbf{E}\|\underline{u}^T\|^2/(T+1) \leq \mathcal{P} - \epsilon$. As ϵ vanishes, those $(\mathbf{K}_T^{(T)}, \mathbf{B}_T)_\epsilon$ can arbitrarily approach $C_T(\mathcal{P})$ or $C_{T,T}(\mathcal{P})$. Now fix $(\mathbf{K}_T^{(T)}, \mathbf{B}_T)_\epsilon$, and construct an admissible sequence $(\mathbf{A}_i, \underline{\mathbf{C}}_i, \mathbf{G}_{T,i})_\epsilon$ as ii) does. Clearly, for sufficiently large i , the input $\underline{u}^T(\mathbf{A}_i, \underline{\mathbf{C}}_i, \mathbf{G}_{T,i})_\epsilon$ satisfies the power constraint $\mathbf{E}\|\underline{u}^T\|^2/(T+1) \leq \mathcal{P}$. That is, for any rate achievable by $\underline{u}^T(\mathbf{K}_T^{(T)}, \mathbf{B}_T)_\epsilon$, it can be arbitrarily approached by $\underline{u}^T(\mathbf{A}_i, \underline{\mathbf{C}}_i, \mathbf{G}_{T,i})_\epsilon$ satisfying the power constraint. Then the result follows.

APPENDIX IV

PROOF OF PROPERTIES OF THE CODING STRUCTURE \mathbb{S}^*

1) *Proof of Proposition 4:* Here we show that the coding structure \mathbb{S}^* , in the form of (62), satisfies the necessary condition for optimality as presented in Proposition 4.

Since $\{y_t\}$ is interchangeable with the innovations process $\{e_t\}$, in the sense that they determine each other causally and linearly, it suffices to show that $\mathbf{E}u_t e_\tau = 0$. Note that

$$u_t = \mathbf{D}\mathbf{X}_t = \mathbf{D}'\mathbf{A}\mathbf{X}_{t-1} - \mathbf{D}'\mathbf{L}_{t-1}e_{t-1},\tag{121}$$

and thus

$$\begin{aligned}\mathbf{E}u_t e_{t-1} &= \mathbf{E}\mathbf{D}'\mathbf{A}\mathbf{X}_{t-1}e_{t-1} - \mathbf{E}\mathbf{D}'\mathbf{L}_{t-1}\mathbf{K}_{e,t-1} \\ &\stackrel{(a)}{=} \mathbf{E}\mathbf{D}'\mathbf{A}\mathbf{X}_{t-1}\mathbf{X}_{t-1}'\mathbf{C} + \mathbf{E}\mathbf{D}'\mathbf{A}\mathbf{X}_{t-1}\mathbf{N}_{t-1} - \mathbf{D}'\mathbf{A}\mathbf{\Sigma}_{t-1}\mathbf{C} \\ &= \mathbf{D}'\mathbf{A}\mathbf{\Sigma}_{t-1}\mathbf{C} + 0 - \mathbf{D}'\mathbf{A}\mathbf{\Sigma}_{t-1}\mathbf{C} = 0,\end{aligned}\tag{122}$$

where (a) follows from (62) and (64). Similarly we can prove $\mathbf{E}u_t e_\tau = 0$ for any $\tau < t-1$.

2) *Proof of Proposition 5:* We first prove a simple technical lemma that is useful in the following development. It says that the product of a banded, lower triangular matrix with bandwidth $(m+1)$ and an upper triangular matrix is banded in its lower triangular part with bandwidth $(m+1)$.

Lemma 4. Suppose $\mathbf{A} \in \mathbb{R}^{n \times n}$ is banded and lower triangular with bandwidth $(m+1)$, i.e., $\mathbf{A}(i, k) = 0$ if $i < k$ or $i > (k+m)$. Suppose $\mathbf{B} \in \mathbb{R}^{n \times n}$ is upper triangular, i.e., $\mathbf{B}(k, j) = 0$ if $k > j$. Then $\mathbf{C} := \mathbf{A}\mathbf{B}$ is banded in its lower triangular part with bandwidth $(m+1)$, i.e., $\mathbf{C}(i, j) = 0$ if $i > j+m$.

Proof: Simply note that $\mathbf{C}(i, j) = \sum_{k=1}^n \mathbf{A}(i, k)\mathbf{B}(k, j)$. So if $i > j+m$, then $\mathbf{A}(i, k)\mathbf{B}(k, j) = 0$ for any k .

□

Now we go back to Proposition 5. To show this proposition, it is sufficient to show that $\mathbf{K}_{\underline{y}}^{(T)}$ is banded in its lower triangular part with bandwidth $(m+1)$, since $\mathbf{K}_{\underline{y}}^{(T)}$ is symmetric. By Lemma 4, we only need to show that $\mathbf{K}_{\underline{y}}^{(T)}$ can be written as the sum of products of banded, lower triangular matrices (of bandwidth $(m+1)$) with upper triangular matrices. To this aim, some algebra shows that

$$\begin{aligned} \mathbf{E}\underline{N}^T \underline{u}^{T'} &= \mathbf{E}\underline{N}^T ((\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1}) \underline{u}^T - \hat{\mathbf{G}}_T \underline{N}^T)' = -\hat{\mathbf{G}}_T' \\ \mathbf{Z}_{p,T} \mathbf{E}\underline{u}^T \underline{y}^{T'} &= \mathbf{Z}_{p,T} \mathbf{K}_{\underline{u}}^T \mathbf{Z}_{p,T}' + \mathbf{Z}_{p,T} \mathbf{E}\underline{u}^T \underline{N}^{T'} \mathbf{Z}_{z,T}' \\ \mathbf{K}_{\underline{y}}^{(T)} &= \mathbf{Z}_{p,T} \mathbf{K}_{\underline{u}}^T \mathbf{Z}_{p,T}' + \mathbf{Z}_{z,T} \mathbf{Z}_{z,T}' + \mathbf{Z}_{p,T} \mathbf{E}\underline{u}^T \underline{N}^{T'} \mathbf{Z}_{z,T}' - \mathbf{Z}_{z,T} \hat{\mathbf{G}}_T' \mathbf{Z}_{p,T}' \\ &= \mathbf{Z}_{p,T} \mathbf{E}\underline{u}^T \underline{y}^{T'} + \mathbf{Z}_{z,T} \mathbf{Z}_{z,T}' - \mathbf{Z}_{z,T} \hat{\mathbf{G}}_T' \mathbf{Z}_{p,T}'. \end{aligned} \quad (123)$$

As $\mathbf{Z}_{z,T}$ is lower triangular, $\mathbf{E}\underline{u}^T \underline{y}^{T'}$ is upper triangular. Therefore, on the right-hand-side of the last equality, $\mathbf{Z}_{p,T}$ and $\mathbf{Z}_{z,T}$ are banded and lower triangular with bandwidth $(m+1)$, and $\mathbf{E}\underline{u}^T \underline{y}^{T'}$, $\mathbf{Z}_{z,T}'$, and $\hat{\mathbf{G}}_T' \mathbf{Z}_{p,T}'$ are upper triangular. Then the result follows.

As an alternative proof or a verification of the above result, let us consider the mapping $\mathcal{M}_{\underline{e},\underline{y}}$ from \underline{e}^T to \underline{y}^T (incorporating the feedback loop). Since

$$\mathbf{K}_{\underline{y}}^{(T)} = \mathcal{M}_{\underline{e},\underline{y}} \mathbf{K}_{\underline{e}}^{(T)} \mathcal{M}_{\underline{e},\underline{y}}' \quad (124)$$

and $\mathbf{K}_{\underline{y}}^{(T)} > 0$, $\mathcal{M}_{\underline{e},\underline{y}}$ is lower triangular and uniquely defined (cf. [23] for relevant discussions of innovation processes, QR factorization, and Cholesky factorization). It is sufficient to show that $\mathcal{M}_{\underline{e},\underline{y}}$ is banded and lower triangular with bandwidth $(m+1)$. To this aim, we characterize $\mathcal{M}_{\underline{e},\underline{y}}$ in state space. Note that the state-space representation from \underline{y}^T to \underline{e}^T is $(\mathbf{F} - \underline{L}_{2,t} \underline{H}', \underline{L}_{2,t}, -\underline{H}', 1)$ and hence the one from \underline{e}^T to \underline{y}^T is $(\mathbf{F}, -\underline{L}_{2,t}, -\underline{H}', 1)$ (see Appendix I-A), and that the state-space representation from \underline{y}^T to \underline{y}^T is $(\mathbf{F}_z, \underline{G}_z, \underline{H}', 1)$ (i.e., this is the state-space for the transfer function $\mathcal{Z}_z(z)$). Hence, we obtain the state-space from \underline{e}^T to \underline{y}^T is

$$\left(\begin{pmatrix} \mathbf{F} & \mathbf{0} \\ -\underline{G}_z \underline{H}' & \mathbf{F}_z \end{pmatrix}, \begin{pmatrix} -\underline{L}_{2,t} \\ \underline{G}_z \end{pmatrix}, (-\underline{H}', \underline{H}', 1) \right); \quad (125)$$

where $\mathbf{F}_z := \mathbf{F} + \underline{G}_z \underline{H}' \in \mathbb{R}^{m \times m}$ is a nilpotent matrix, that is, $\mathbf{F}_z^m = \mathbf{0}$ and only the $(1, m)$ th entry of \mathbf{F}_z^{m-1} is non-zero (equal to 1). The above state-space realization is not minimal. Simple computation shows that the above $2m$ th order representation is equivalent to the m th order representation $(\mathbf{F}_z, \underline{G}_{e,t}, \underline{H}', 1)$ where $\underline{G}_{e,t} := \underline{G}_z + \underline{L}_{2,t}$. From the relation between the state-space representation and the impulse response, it holds that the (i, j) th entry in $\mathcal{M}_{\underline{e},\underline{y}}$ is $\underline{H}' \mathbf{F}_z^{i-j-1} \underline{G}_{e,j-1}$, if $i > j$. Notice that $\mathcal{M}_{\underline{e},\underline{y}}(i, i) = 1$. Because \mathbf{F}_z is nilpotent matrix, we have $\mathcal{M}_{\underline{e},\underline{y}}(i, j) = 0$ if $(i - j - 1) \geq m$, namely, the lower triangular part of $\mathcal{M}_{\underline{e},\underline{y}}$ is banded with bandwidth $(m+1)$. Thus, $\mathbf{K}_{\underline{y}}^{(T)}$ is banded with bandwidth $(m+1)$.

A. Proof of Proposition 6

We follow the following steps to prove the lemma. First, by considering the equivalent open loop, non-feedback communication problem, we show that in order for an \underline{u}^T to achieve $C_{T,n}$, it must hold that the effective channel input $(\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1}) \underline{u}^T$ waterfills the effective channel noise $(\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1})^{-1} \mathbf{Z}_T \underline{N}^T$. Second, we show this would yield that, if the optimizing $\mathbf{K}_{\underline{r}}^{(T)}$ has rank k , then the optimal channel output covariance matrix $\mathbf{K}_{\underline{y}}^{(T)}$ has its smallest (positive) eigenvalue, denoted λ_0 , repeated exactly k times. Therefore, $(\mathbf{K}_{\underline{y}}^{(T)} - \lambda_0 \mathbf{I})$ has rank $(T+1-k)$. This in turn results in that

$$\mathbf{M} := \mathbf{K}_{\underline{y}}^{(T)} - \lambda_0 \mathbf{Z}_{p,T} \mathbf{Z}_{p,T}' \quad (126)$$

has rank $(T+1-k)$. However, since $\mathbf{K}_{\underline{y}}^{(T)}$ and $\mathbf{Z}_{p,T} \mathbf{Z}_{p,T}'$ are banded with bandwidth $(2m+1)$, it can be shown that \mathbf{M} has rank at least $(T+1-m)$. So k has to be no larger than m . The details follow.

First, we prove that the effect channel input needs to waterfill the effective channel noise to achieve $C_{T,n}$. Since

$$\tilde{\mathbf{y}}^T = \mathbf{Z}_T \mathbf{y}^T = (\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1})^{-1} (\mathbf{r}^T + \mathbf{Z}_T \mathbf{N}^T) \quad (127)$$

and

$$\mathbf{K}_{\tilde{\mathbf{y}}}^{(T)} = (\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1})^{-1} \mathbf{K}_{\mathbf{r}}^{(T)} (\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1})^{-1'} + (\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1})^{-1} \mathbf{G}_T \mathbf{G}_T' (\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1})^{-1'} := \mathbf{P}_r + \mathbf{P}_Z \quad (128)$$

where \mathbf{P}_r and \mathbf{P}_Z are defined in obvious way, the optimization problem for $C_{T,n}(\mathcal{P})$ can be recast as

$$C_{T,n}(\mathcal{P}) = \sup_{\substack{\mathbf{A} \in \mathbb{R}^{(n+1) \times (n+1)}, \underline{\mathbf{G}}, \mathbf{G}_T \\ s.t. \text{ trace}(\mathbf{P}_r + \mathbf{P}_Z)/(T+1) \leq \mathcal{P}}} \frac{1}{T+1} \log \det(\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1})^{-1} (\mathbf{K}_{\mathbf{r}}^{(T)} + \mathbf{Z}_T \mathbf{Z}_T') (\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1})^{-1'}. \quad (129)$$

If the optimizing \mathbf{G}_T^* is plugged into the above optimization problem, noticing the resulting \mathbf{P}_Z is independent of the choice of $\mathbf{K}_{\mathbf{r}}^{(T)}$, we end up with the following optimization problem

$$C_{T,n}(\mathcal{P}) := \sup_{\substack{\mathbf{A} \in \mathbb{R}^{(n+1) \times (n+1)}, \underline{\mathbf{G}}, \mathbf{G}_T, (A1) \\ s.t. \text{ trace}(\mathbf{P}_r)/(T+1) \leq \mathcal{P}}} \frac{1}{2(T+1)} \log \det(\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1})^{-1} (\mathbf{K}_{\mathbf{r}}^{(T)} + \mathbf{Z}_T \mathbf{Z}_T') (\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1})^{-1'}. \quad (130)$$

This may be viewed as a finite-horizon *non-feedback* capacity problem, in which the effective input to the channel without feedback is $(\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1})^{-1} \mathbf{r}^T$, and the effective channel noise is $(\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1})^{-1} \mathbf{Z}_T \mathbf{N}^T$. This idea has been used in [33], [45], [51]. Thus, in order to give rise to the maximum mutual information between the effective input and the channel output $\tilde{\mathbf{y}}^T$, it is necessary to have the effective input to waterfill the effective noise.

The implication of the waterfilling argument is that, if the (effective) input covariance matrix has rank k , then the output covariance matrix must have its smallest eigenvalue repeated exactly k times. This is intuitively suggested by the name “waterfilling”. In other words, if the waterfilling level (cutoff value) is λ_0 , then there are k eigenvalues of $\mathbf{K}_{\tilde{\mathbf{y}}}^{(T)}$ that are equal to λ_0 and are the consequence of “waterfilling” by k positive eigenvalues of $(\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1})^{-1} \mathbf{K}_{\mathbf{r}}^{(T)} (\mathbf{I} - \mathbf{G}_T \mathbf{Z}_T^{-1})^{-1'}$ or equivalently $\mathbf{K}_{\mathbf{r}}^{(T)}$, and the rest of the eigenvalues are strictly larger than λ_0 that remain unchanged after the waterfilling. Hence, if the optimizing $\mathbf{K}_{\mathbf{r}}^{(T)}$ has rank k , then the resulting $(\mathbf{K}_{\tilde{\mathbf{y}}}^{(T)} - \lambda_0 \mathbf{I})$ has rank $(T+1-k)$. Since $\mathbf{Z}_{p,T}$ has full rank, it holds that

$$\mathbf{M} := \mathbf{K}_{\tilde{\mathbf{y}}}^{(T)} - \lambda_0 \mathbf{Z}_{p,T} \mathbf{Z}_{p,T}' = \mathbf{Z}_{p,T} (\mathbf{K}_{\tilde{\mathbf{y}}}^{(T)} - \lambda_0 \mathbf{I}) \mathbf{Z}_{p,T}' \quad (131)$$

has rank $(T+1-k)$. Since $\mathbf{K}_{\tilde{\mathbf{y}}}^{(T)}$ and $\mathbf{Z}_{p,T} \mathbf{Z}_{p,T}'$ are banded with bandwidth no larger than $(2m+1)$, we have that \mathbf{M} is banded with bandwidth $(2m+1)$. By $(f_0 + g_0) = 0$, it holds that $\mathbf{Z}_{p,T}(j+m, j) = 0$ for any j . Therefore, $\mathbf{M}(j+m, j) = \mathbf{K}_{\tilde{\mathbf{y}}(T)}(j+m, j)$. However, from the proof of Proposition 5, we have

$$\mathbf{K}_{\tilde{\mathbf{y}}}^{(T)} = \mathbf{Z}_{p,T} \mathbf{E} \mathbf{u}^T \tilde{\mathbf{y}}^{T'} + \mathbf{Z}_{z,T} \mathbf{Z}_{z,T}' - \mathbf{Z}_{z,T} \hat{\mathbf{G}}_T' \mathbf{Z}_{p,T}', \quad (132)$$

which leads to that $\mathbf{K}_{\tilde{\mathbf{y}}(T)}(j+m, j) = f_0 \neq 0$ for any j ; notice that $\hat{\mathbf{G}}_T$ is strictly lower triangular and $\mathbf{Z}_{p,T}(j+m, j) = 0$. Then the banded structure of \mathbf{M} implies that the rank of \mathbf{M} is at least $(T+1-m)$. This immediately follows that k has to be no larger than m . This proves Lemma 2.

Now we go back to Proposition 6. For any m th order channel $\mathcal{Z}(z)$, consider the following perturbation that leads to an $(m+1)$ st order channel:

$$\begin{aligned} \mathcal{Z}_\epsilon(z) &:= \frac{\mathcal{Z}_{z,\epsilon}(z)}{\mathcal{Z}_p(z)} \\ &:= \frac{(1-\epsilon)(1+f_{m-1}z^{-1}+\dots+f_1z^{-m+1}+f_0z^{-m})-\epsilon^iz^{-m-1}}{1+(f_{m-1}+g_{m-1})z^{-1}+\dots+(f_1+g_1)z^{-m+1}+(f_0+g_0)z^{-m}}, \end{aligned} \quad (133)$$

where $\epsilon > 0$ and i is an integer to be determined. In other words, $\mathcal{Z}_z(z)$ is perturbed to be $\mathcal{Z}_{z,\epsilon}(z)$. Consequently, $\mathcal{Z}_{p,\epsilon,T} = (1 - \epsilon)\mathcal{Z}_{p,T} - \epsilon^i \mathbf{J}$, where \mathbf{J} is the down-shift matrix with $\mathbf{J}(j + m + 1, j) = 1$ for any j and all other entries equal to 0.

The number i is chosen such that for sufficiently small $\epsilon > 0$, we have $\mathcal{Z}_{z,T}\mathcal{Z}'_{z,T} > \mathcal{Z}_{z,\epsilon}\mathcal{Z}'_{z,\epsilon}$. Such an i always exists. To see this, note that the difference of the two covariance matrices is

$$\begin{aligned} & \epsilon(2 - \epsilon)\mathcal{Z}_{z,T}\mathcal{Z}'_{z,T} + \epsilon^i(1 - \epsilon)(\mathbf{J}\mathcal{Z}'_{z,T} + \mathcal{Z}_{z,T}\mathbf{J}') - \epsilon^{2i}\mathbf{J}\mathbf{J}' \\ & > \epsilon(2 - \epsilon)\mathcal{Z}_{z,T}\mathcal{Z}'_{z,T} + \epsilon^i(1 - \epsilon)(\mathbf{J}\mathcal{Z}'_{z,T} + \mathcal{Z}_{z,T}\mathbf{J}' - \mathbf{J}\mathbf{J}'). \end{aligned} \quad (134)$$

As $\mathcal{Z}_{z,T}\mathcal{Z}'_{z,T}$ is positive definite and $(\mathbf{J}\mathcal{Z}'_{z,T} + \mathcal{Z}_{z,T}\mathbf{J}' - \mathbf{J}\mathbf{J}')$ is symmetric, the above difference admits simultaneous diagonalization, which transforms both terms to be diagonal. That is, there exists \mathbf{S} non-singular such that the congruence transformation using \mathbf{S} leads to

$$\epsilon(2 - \epsilon)\mathbf{D}_1 + \epsilon^i(1 - \epsilon)\mathbf{D}_2 \quad (135)$$

with both \mathbf{D}_1 and \mathbf{D}_2 are diagonal and independent of ϵ . Suppose $\rho := \max_j |\mathbf{D}_2(j, j)/\mathbf{D}_1(j, j)|$. Then for any i such that

$$i > 1 + \frac{\log \frac{2-\epsilon}{\rho(1-\epsilon)}}{\log \epsilon}, \quad (136)$$

the above difference is positive definite. However, for sufficiently small ϵ , the right-hand-side of the inequality approaches 1, so it is sufficient to choose $i := 2$, independent of ϵ . Hence $\mathbf{K}_{\mathcal{Z}}^{(T)} > \mathbf{K}_{\mathcal{Z}_\epsilon}^{(T)}$. Similarly we can show

$$\mathbf{K}_{\mathcal{Z}_{\epsilon_1}}^{(T)} > \mathbf{K}_{\mathcal{Z}_{\epsilon_2}}^{(T)} \quad (137)$$

if $\epsilon_1 < \epsilon_2$.

Next we show that as $\epsilon > 0$ approaches zero, $P_T(\mathcal{R}, \mathcal{Z}_{\epsilon,T})$ admits a limit which is no larger than $P_T(\mathcal{R}, \mathcal{Z}_T)$. Due to $\mathbf{K}_{\mathcal{Z}}^{(T)} \geq \mathbf{K}_{\mathcal{Z}_\epsilon}^{(T)}$, the feasible set $\phi_{\mathcal{Z},T,\mathcal{R}}$ for \mathcal{Z}_T is strictly contained in $\phi_{\mathcal{Z}_\epsilon,T,\mathcal{R}}$. Using the ordering in (137) and the capacity inequality proven in [15], it is seen that $P_T(\mathcal{R}, \mathcal{Z}_{\epsilon,T})$ is no larger than $P_T(\mathcal{R}, \mathcal{Z}_T)$ and is non-decreasing as ϵ approaches zero from above and hence the limit exists.

Consider an arbitrarily small slack $\delta > 0$ of rate, i.e., consider $P_T(\mathcal{R} - \delta, \mathcal{Z}_T)$. For any $\delta > 0$, there exist $\epsilon > 0$ such that the feasible set $\phi_{\mathcal{Z}_\epsilon,T,\mathcal{R}}$ is contained in $\phi_{\mathcal{Z}_\epsilon,T,\mathcal{R}-\delta}$. To see such an ϵ always exists, note that

$$\det(I + \mathbf{\Gamma}_T(\mathbf{K}_{\mathcal{Z}_\epsilon}^{(T)})^{-1}\mathbf{\Gamma}'_T) = \det(I + \mathbf{\Gamma}_T(\mathbf{K}_{\mathcal{Z}_\epsilon}^{(T)})^{-1}\mathbf{\Gamma}'_T + \Delta_\epsilon), \quad (138)$$

where $\Delta_\epsilon \geq 0$ and vanishes as ϵ tends to zero. Thus if

$$\frac{1}{2} \log \det(I + \mathbf{\Gamma}_T(\mathbf{K}_{\mathcal{Z}_\epsilon}^{(T)})^{-1}\mathbf{\Gamma}'_T) \geq \mathcal{R} \quad (139)$$

then

$$\frac{1}{2} \log \det(I + \mathbf{\Gamma}_T(\mathbf{K}_{\mathcal{Z}_\epsilon}^{(T)})^{-1}\mathbf{\Gamma}'_T) \geq \mathcal{R} - \delta \quad (140)$$

for small enough ϵ .

Interconnecting the optimizing $\mathcal{r}_{\epsilon,T}$ and $\mathcal{G}_{\epsilon,T}$ (obtained for $P_T(\mathcal{R}, \mathcal{Z}_{\epsilon,T})$)¹⁷ with \mathcal{Z}_T , we see that the interconnection satisfies the rate constraint for $P_T(\mathcal{R} - \delta, \mathcal{Z}_T)$. The consumed power becomes

$$\text{trace}[(I - \mathcal{G}_{\epsilon,T}\mathcal{Z}_{\epsilon,T}^{-1})^{-1}(\mathbf{K}_{\mathcal{Z}_\epsilon}^{(T)} + \mathcal{G}_{\epsilon,T}\mathbf{K}_{\mathcal{Z}}^{(T)}\mathcal{G}'_{\epsilon,T})(I - \mathcal{G}_{\epsilon,T}\mathcal{Z}_{\epsilon,T}^{-1})], \quad (141)$$

¹⁷To be more rigorous, an arbitrarily small slack to $P_T(\mathcal{R}, \mathcal{Z}_{\epsilon,T})$ may be needed since this optimization problem is an infimization problem as opposed to a minimization problem. The idea in the proof can be easily adopted when the slack is used and the same result holds.

which is greater than

$$P_T(\mathcal{R}, \mathcal{Z}_{\epsilon, T}) = \text{trace}[(\mathbf{I} - \mathcal{G}_{\epsilon, T} \mathcal{Z}_{\epsilon, T}^{-1})^{-1} (\mathbf{K}_{\mathcal{Z}, \epsilon}^{(T)} + \mathcal{G}_{\epsilon, T} \mathbf{K}_{\mathcal{Z}, \epsilon}^{(T)} \mathcal{G}_{\epsilon, T}' (\mathbf{I} - \mathcal{G}_{\epsilon, T} \mathcal{Z}_{\epsilon, T}^{-1})'] \quad (142)$$

but the difference dependent on $(\mathbf{K}_{\mathcal{Z}}^{(T)} - \mathbf{K}_{\mathcal{Z}, \epsilon}^{(T)})$ vanishes as ϵ goes to zero. Consequently, as ϵ goes to zero, the consumed power of the sequence of interconnections converges to $P_T(\mathcal{R}, \mathcal{Z}_{\epsilon, T})$, no larger than $P_T(\mathcal{R}, \mathcal{Z}_T)$. However, each optimizing $\mathbf{K}_{\mathcal{Z}, \epsilon}^{(T)}$ has rank no larger than $(m+1)$ by Lemma 2. Therefore, $P_T(\mathcal{R}, \mathcal{Z}_T)$ can be achieved by a sequence of $\mathbf{K}_{\mathcal{Z}, \epsilon}^{(T)}$ with rank no larger than $(m+1)$. Thus we have proved the proposition.

APPENDIX V

PROOF OF PROPOSITION 8: CONVERGENCE TO STEADY-STATE

We show that system (62) converges to a steady-state, as given by (92). To this aim, we first transform the Riccati recursion into a new coordinate system, then show that it converges to a limit, and finally prove that the limit is the unique stabilizing solution of the Riccati equation. The convergence to the steady-state follows immediately from the convergence of the Riccati recursion.

Consider a coordinate transformation given as

$$\tilde{\mathbf{A}} := \Psi \mathbf{A} \Psi^{-1} := \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{F} \end{bmatrix}, \quad \tilde{\mathbf{C}} := \Psi^{-1'} \mathbf{C} = \begin{bmatrix} \mathbf{C} + \phi' \mathbf{H} \\ \mathbf{H} \end{bmatrix}, \quad \tilde{\mathbf{D}} := \Psi^{-1'} \mathbf{D} = \begin{bmatrix} \mathbf{C} \\ \mathbf{0} \end{bmatrix}, \quad \tilde{\Sigma}_t := \Psi \Sigma_t \Psi', \quad (143)$$

where

$$\Psi := \begin{bmatrix} \mathbf{I}_{n+1} & \mathbf{0} \\ -\psi & \mathbf{I}_m \end{bmatrix} \quad (\text{i.e. } \Psi^{-1} = \begin{bmatrix} \mathbf{I}_{n+1} & \mathbf{0} \\ \psi & \mathbf{I}_m \end{bmatrix}), \quad (144)$$

and ψ is the unique solution to the Sylvester equation

$$\mathbf{F}\psi - \psi\mathbf{A} = -\mathbf{G}\mathbf{C}'. \quad (145)$$

Note that the existence and uniqueness of ψ is guaranteed by the assumption on \mathbf{A} that $\lambda_i(-\mathbf{A}) + \lambda_j(\mathbf{F}) \neq 0$ for any i and j (see Section V-A).

Assume $k = n$ for the rest of the proof; i.e., \mathbf{A} is anti-stable. For the case $k < n$, we can further transform $\tilde{\mathbf{A}}$, $\tilde{\mathbf{C}}$, and $\tilde{\Sigma}$ into $\tilde{\tilde{\mathbf{A}}}$, $\tilde{\tilde{\mathbf{C}}}$, and $\tilde{\tilde{\Sigma}}$ such that

$$\tilde{\tilde{\mathbf{A}}} = \text{diag}[\mathbf{A}_+, \mathbf{A}_-, \mathbf{F}], \quad (146)$$

where $\mathbf{A}_+ \in \mathbb{R}^{(k+1) \times (k+1)}$ is anti-stable and \mathbf{A}_- is stable; then the following argument can be easily modified for the case $k < n$.

we can further decompose \mathbf{A} into block-diagonal form and incorporate the stable block of \mathbf{A} into \mathbf{F} , and the proof follows similarly; note that $\tilde{\tilde{\mathbf{A}}}$ does not have any eigenvalues on the unit circle.

The transformation defined in (143) transforms \mathbf{A} into a block-diagonal form with the unstable and stable eigenvalues in different on-diagonal blocks, and transforms the initial condition Σ_0 to

$$\tilde{\Sigma}_0 := \Psi \begin{bmatrix} \mathbf{I}_{n+1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \Psi' = \begin{bmatrix} \mathbf{I} & -\psi' \\ -\psi & \psi\psi' \end{bmatrix}. \quad (147)$$

Therefore, the convergence of (65) with initial condition Σ_0 is equivalent to the convergence of

$$\tilde{\Sigma}_{t+1} = \tilde{\tilde{\mathbf{A}}} \tilde{\Sigma}_t \tilde{\tilde{\mathbf{A}}} - \frac{\tilde{\tilde{\mathbf{A}}} \tilde{\Sigma}_t \tilde{\tilde{\mathbf{C}}} \tilde{\tilde{\mathbf{C}}} \tilde{\Sigma}_t \tilde{\tilde{\mathbf{A}}}}{\tilde{\tilde{\mathbf{C}}} \tilde{\Sigma}_t \tilde{\tilde{\mathbf{C}}} + 1} \quad (148)$$

with initial condition $\tilde{\Sigma}_0$. By [47], $\tilde{\Sigma}_t$ would converge if

$$\det \left(\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_m \end{bmatrix} - \tilde{\Sigma}_0 \begin{bmatrix} \mathbf{I}_{n+1} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{22} \end{bmatrix} \right) \neq 0, \quad (149)$$

where \mathbf{X}_{22} is the negative semi-definite matrix to the discrete-time Lyapunov equation

$$\mathbf{X}_{22} = \mathbf{F}\mathbf{X}_{22}\mathbf{F}' - (\underline{\mathbf{C}} + \underline{\boldsymbol{\psi}}'\underline{\mathbf{H}})(\underline{\mathbf{C}} + \underline{\boldsymbol{\psi}}'\underline{\mathbf{H}})'. \quad (150)$$

Notice that $(\underline{\mathbf{C}} + \underline{\boldsymbol{\psi}}'\underline{\mathbf{H}})$ is the upper $(n+1) \times 1$ block in $\tilde{\underline{\mathbf{C}}}$. Since

$$\begin{aligned} \det \left(\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} - \begin{bmatrix} \mathbf{I} & -\underline{\boldsymbol{\psi}}' \\ -\underline{\boldsymbol{\psi}} & \underline{\boldsymbol{\psi}}\underline{\boldsymbol{\psi}}' \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{22} \end{bmatrix} \right) &= \det \left(\begin{bmatrix} -\mathbf{I} & \underline{\boldsymbol{\psi}}'\mathbf{X}_{22} \\ \underline{\boldsymbol{\psi}} & \mathbf{I} - \underline{\boldsymbol{\psi}}\underline{\boldsymbol{\psi}}'\mathbf{X}_{22} \end{bmatrix} \right) \\ &= \det(-\mathbf{I}) \det(\mathbf{I} - \underline{\boldsymbol{\psi}}\underline{\boldsymbol{\psi}}'\mathbf{X}_{22} + \underline{\boldsymbol{\psi}}\underline{\boldsymbol{\psi}}'\mathbf{X}_{22}) \\ &\neq 0, \end{aligned} \quad (151)$$

we conclude that $\tilde{\boldsymbol{\Sigma}}_t$ converges to a limit $\tilde{\boldsymbol{\Sigma}}_\infty$.

This limit $\tilde{\boldsymbol{\Sigma}}_\infty$ is a positive semi-definite solution to

$$\tilde{\boldsymbol{\Sigma}}_\infty = \tilde{\mathbf{A}}\tilde{\boldsymbol{\Sigma}}_\infty\tilde{\mathbf{A}}' - \frac{\tilde{\mathbf{A}}\tilde{\boldsymbol{\Sigma}}_\infty\tilde{\underline{\mathbf{C}}}\tilde{\underline{\mathbf{C}}}'\tilde{\boldsymbol{\Sigma}}_\infty\tilde{\mathbf{A}}'}{\tilde{\underline{\mathbf{C}}}'\tilde{\boldsymbol{\Sigma}}_\infty\tilde{\underline{\mathbf{C}}} + 1}. \quad (152)$$

By [23], (152) has a unique stabilizing solution because $(\tilde{\mathbf{A}}, \tilde{\underline{\mathbf{C}}})$ is observable (noting that $(\mathbf{A}, \underline{\mathbf{C}}')$ is observable) and $\tilde{\mathbf{A}}$ does not have any eigenvalues on the unit circle. Therefore, $\tilde{\boldsymbol{\Sigma}}_\infty$ is this unique stabilizing solution, which can be computed from (152) as (see also [47])

$$\tilde{\boldsymbol{\Sigma}}_\infty = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad (153)$$

where $\tilde{\boldsymbol{\Sigma}}_{11}$ is the positive-definite solution to a reduced-order Riccati equation

$$\tilde{\boldsymbol{\Sigma}}_{11} = \mathbf{A}\tilde{\boldsymbol{\Sigma}}_{11}\mathbf{A}' - \frac{\mathbf{A}\tilde{\boldsymbol{\Sigma}}_{11}(\underline{\mathbf{C}}' + \underline{\mathbf{H}}'\underline{\boldsymbol{\psi}})'(\underline{\mathbf{C}}' + \underline{\mathbf{H}}'\underline{\boldsymbol{\psi}})\tilde{\boldsymbol{\Sigma}}_{11}\mathbf{A}'}{(\underline{\mathbf{C}}' + \underline{\mathbf{H}}'\underline{\boldsymbol{\psi}})\tilde{\boldsymbol{\Sigma}}_{11}(\underline{\mathbf{C}}' + \underline{\mathbf{H}}'\underline{\boldsymbol{\psi}})' + 1}. \quad (154)$$

and has rank equal to the number of anti-stable eigenvalues of $\tilde{\mathbf{A}}$ (cf. [47]). Thus, $\boldsymbol{\Sigma}_t$ converges to

$$\boldsymbol{\Sigma}_\infty = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{11} & \tilde{\boldsymbol{\Sigma}}_{11}\underline{\boldsymbol{\psi}}' \\ \underline{\boldsymbol{\psi}}\tilde{\boldsymbol{\Sigma}}_{11} & \underline{\boldsymbol{\psi}}\tilde{\boldsymbol{\Sigma}}_{11}\underline{\boldsymbol{\psi}}' \end{bmatrix} \quad (155)$$

with rank equal to the number of anti-stable eigenvalues of $\tilde{\mathbf{A}}$.

ii) Immediately from i).

APPENDIX VI

PROOF OF COROLLARY 1

Consider the coordinate transformations used in the proof of Proposition 8 that transform \mathbf{A} , $\underline{\mathbf{C}}$, and $\boldsymbol{\Sigma}$ into $\tilde{\mathbf{A}}$, $\tilde{\underline{\mathbf{C}}}$, and $\tilde{\boldsymbol{\Sigma}}$. Note that the block in $\tilde{\boldsymbol{\Sigma}}_\infty$ (i.e. the solution to the Riccati equation defined by $\tilde{\mathbf{A}}$ and $\tilde{\underline{\mathbf{C}}}$) associated with the \mathbf{A}_- block is zero. By Proposition 9, in the new coordinates the rate and power due to the \mathbf{A}_- block are both zero, and hence in the original coordinates the rate and power due to the stable eigenvalues of \mathbf{A} are both zero. Then we remove the dimensions corresponding to \mathbf{A}_- in $\tilde{\mathbf{A}}$, $\tilde{\underline{\mathbf{C}}}$, $\tilde{\boldsymbol{\Sigma}}$, and the coordinate transformation matrix. It is easy to check that this leads to a pair of reduced order $(\mathbf{A}_k, \underline{\mathbf{C}}_k)$ with \mathbf{A}_k anti-stable and satisfying $R_{\infty,n}(\mathbf{A}, \underline{\mathbf{C}}) = R_{\infty,k}(\mathbf{A}_k, \underline{\mathbf{C}}_k)$ and $P_{\infty,n}(\mathbf{A}, \underline{\mathbf{C}}) = P_{\infty,k}(\mathbf{A}_k, \underline{\mathbf{C}}_k)$.

ACKNOWLEDGEMENTS

The authors would like to thank Anant Sahai, Sekhar Tatikonda, Sanjoy Mitter, Zhengdao Wang, Murti Salapaka, Shaohua Yang, Donatello Materassi, and Young-Han Kim for useful discussion.

REFERENCES

- [1] J. P. M. Schalkwijk and T. Kailath. A coding scheme for additive noise channels with feedback Part I: No bandwidth constraint. *IEEE Trans. Inform. Theory*, IT-12(2):172–182, Apr. 1966.
- [2] J. P. M. Schalkwijk. A coding scheme for additive noise channels with feedback? Part II: Bandlimited signals. *IEEE Trans. Inform. Theory*, IT-12(2):183–189, Apr. 1966.
- [3] J. K. Omura. Optimum linear transmission of analog data for channels with feedback. *IEEE Trans. Inform. Theory*, 14(1):38–43, Jan. 1968.
- [4] S. A. Butman. A general formulation of linear feedback communication systems with solutions. *IEEE Trans. Inform. Theory*, IT-15:392–400, 1969.
- [5] S. A. Butman. Linear feedback rate bounds for regressive channels. *IEEE Trans. Inform. Theory*, IT-22:363–366, 1976.
- [6] T. M. Cover and S. Pombra. Gaussian feedback capacity. *IEEE Trans. Inform. Theory*, IT-35:37–43, 1989.
- [7] L. H. Ozarow. Random coding for additive Gaussian channels with feedback. *IEEE Trans. Inform. Theory*, 36(1):17–22, Jan. 1988.
- [8] K. Yanagi. Necessary and sufficient condition for capacity of the discrete-time Gaussian channel to be increased by feedback. *IEEE Trans. Inform. Theory*, 38:1788–1791, Nov. 1992.
- [9] E. Ordentlich. A class of optimal coding schemes for moving average additive Gaussian noise channels with feedback. *Proc. IEEE International Symposium on Information Theory (ISIT)*, page 467, June 1994.
- [10] A. Shahar-Doron and M. Feder. On a capacity achieving scheme for the colored Gaussian channel with feedback. *Proc. 2004 IEEE International Symposium on Information Theory (ISIT)*, page 74, July 2004.
- [11] S. Tatikonda and S. Mitter. The capacity of channels with feedback. *IEEE Trans. Inform. Theory*, 55(1):323–349, Jan. 2009.
- [12] S. Yang, A. Kavcic, and S. Tatikonda. On the feedback capacity of power-constrained Gaussian noise channels with memory. *IEEE Trans. Inform. Theory*, 53(3):929 – 954, Mar. 2007.
- [13] A. Sahai. *Anytime Information Theory*. PhD thesis, MIT, Cambridge, MA, 2001.
- [14] N. Elia. When Bode meets Shannon: Control-oriented feedback communication schemes. *IEEE Trans. Automat. Contr.*, 49(9):1477–1488, Sept. 2004.
- [15] Y. Kim. Feedback capacity of the first-order moving average Gaussian channel. *IEEE Trans. Inform. Theory*, 52(7):3063–3079, July 2006.
- [16] S. Ihara. *Information Theory for Continuous Systems*. World Scientific, 1993.
- [17] J. Massey. Causality, feedback, and directed information. *Proc. IEEE International Symposium on Information Theory and Applications (ISITA)*, pages 303–305, 1990.
- [18] S. Tatikonda. *Control Under Communication Constraints*. PhD thesis, MIT, Cambridge, MA, Aug. 2000.
- [19] L. Vandenbergh, S. Boyd, and S.-P. Wu. Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications*, 19(2):499–533, 1998.
- [20] A. Sahai and S. Mitter. The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link: Parts I: Scalar systems. *IEEE Trans. Inform. Theory*, 52(8):3369–3395, Aug. 2006.
- [21] J. Liu, N. Elia, and S. Tatikonda. Capacity-achieving feedback scheme for Markov channels with channel state information. *Submitted to IEEE Trans. Inform. Theory*, Aug. 2005. See also: Capacity-achieving Feedback Communication Scheme for Markov Channels with Channel State Information. *Proc. 2004 IEEE International Symposium on Information Theory (ISIT)*, page 71, June 2004.
- [22] J. Liu. *Fundamental Limits in Gaussian Channels with Feedback: Confluence of Communication, Estimation, and Control*. PhD thesis, Iowa State University, Ames, IA, Apr. 2006. Available at <http://archives.ece.iastate.edu/archive/00000219/>. See also <http://archives.ece.iastate.edu/archive/00000200/>.
- [23] T. Kailath, A. Sayed, and B. Hassibi. *Linear Estimation*. Prentice Hall, 2000, Englewood Cliffs, NJ, 2000.
- [24] H. L. Van Trees. *Detection, Estimation, and Modulation Theory, Part I*. John Wiley and Sons, New York, 1968.
- [25] D. Guo, S. Shamai, and S. Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Trans. Inform. Theory*, 51(4):1261–1282, Apr. 2005.
- [26] X. Feng, K. A. Loparo, and Y. Fang. Optimal state estimation for stochastic systems: An information theoretic approach. *IEEE Trans. Automat. Contr.*, 42(6):771–785, June 1997.
- [27] P. A. Iglesias. Tradeoffs in linear time-varying systems: an analogue of Bode’s sensitivity integral. *Automatica*, 37:1541–1550, 2001.

- [28] G. Zang and P. A. Iglesias. Nonlinear extension of Bodes integral based on an information theoretic interpretation. *Syst. Control Lett.*, 50:11–19, 2003.
- [29] M. M. Seron, J. H. Braslavsky, and G. C. Goodwin. *Fundamental Limitations in Filtering and Control*. Springer, London, U.K., 2nd ed. edition, 1997.
- [30] N. C. Martins and M. A. Dahleh. Fundamental limitations of disturbance attenuation in the presence of side information. *IEEE Trans. Automat. Contr.*, 52(1):56–66, Jan. 2007.
- [31] N. C. Martins and M. A. Dahleh. Feedback control in the presence of noisy channels: Bode-like fundamental limitations of performance. *IEEE Trans. Automat. Contr.*, 53(7):1604–1615, Aug. 2008.
- [32] Y. Kim. On the feedback capacity of stationary Gaussian channels. *Proc. 43rd Annual Allerton Conference on Communication, Control, and Computing*, Sept. 2005. Available at <http://arxiv.org/abs/cs/0509078>.
- [33] Y. Kim. Feedback capacity of stationary Gaussian channels. *Submitted to IEEE Trans. Inform. Theory*, Feb. 2006. Available at <http://arxiv.org/abs/cs/0602091>.
- [34] R. G. Gallager and Baris Nakiboglu. Variations on a theme by Schalkwijk and Kailath. <http://arxiv.org/abs/0812.2709>, 2009.
- [35] A. Papoulis and S. U. Pillai. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, Boston, MA, 4th edition, 2002.
- [36] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.
- [37] J. P. M. Schalkwijk. Center of gravity information feedback. *IEEE Trans. Inform. Theory*, IT-14:324–331, 1968.
- [38] S. Ihara. Capacity of discrete time Gaussian channel with and without feedback - I. *Mem. Fac. Sci. Kochi Univ. (Math.)*, 9:21–36, 1988.
- [39] S. Ihara. Capacity of discrete time Gaussian channel with and without feedback - II. *Japan Journal of Applied Mathematics*, 6(2):245–258, 1989.
- [40] F. R. Kschischang, B. J. Frey, and H. A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Inform. Theory*, 47(2):498 – 519, Feb. 2001.
- [41] G. D. Forney. On the role of MMSE estimation in approaching the information theoretic limits of linear Gaussian channels: Shannon meets Wiener. *Proc. 41st Annual Allerton Conference on Communication, Control, and Computing*, Oct. 2003.
- [42] S. K. Mitter and N. Newton. Information and entropy flow in the Kalman-Bucy filter. *J. of Stat. Phys.*, 118:145–176, Jan. 2005.
- [43] S. M. Kay. *Fundamentals of Statistical Signal Processing I: Estimation Theory*. Prentice-Hall PTR, Englewood Cliffs, N.J., 1998.
- [44] H. Kwakernaak and R. Sivan. *Linear Optimal Control Systems*. John Wiley & Sons, New York, 1972.
- [45] J. Liu and N. Elia. Achieving the stationary feedback capacity for Gaussian channels. *Proc. 43rd Annual Allerton Conf. on Communication, Control, and Computing*, Sept. 2005. See also <http://arxiv.org/abs/cs.IT/0512097>.
- [46] D. G. Luenberger. *Optimization by Vector Space Methods*. John Wiley and Sons, Inc., 1969.
- [47] K. Gallivan, X. Rao, and P. Van Dooren. Singular Riccati equations stabilizing large-scale systems. *Lin. Alg. Appl.*, 415:359–372, 2006.
- [48] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab. *Signals and Systems*. Prentice Hall, New Jersey, 2nd edition, 1996.
- [49] C. T. Chen. *Linear Systems Theory and Design*. Oxford University Press, New York, 3rd ed. edition, 1999.
- [50] M. A. Dahleh and I. J. Diaz-bobillo. *Control of Uncertain Systems: A Linear Programming Approach*. Prentice Hall, 1995.
- [51] T. E. Klein. *Capacity of Gaussian Noise Channels with Side Information and Feedback*. PhD thesis, MIT, Cambridge, MA, Feb. 2001.